CORTEX XXX (XXXX) XXX

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex

Registered Report

Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow

COLA consortium, Adam J. Parker [a,1,2], Zoe V.J. Woodhead [a,1], David P. Carey [b], Margriet A. Groen [c], Eva Gutierrez-Sigut [d,e], Jessica Hodgson [f], John Hudson [f], Emma M. Karlsson [b], Mairéad MacSweeney [e], Heather Payne [e], Nuala Simpson [a], Paul A. Thompson [a], Kate E. Watkins [a], Ciara Egan [a,b], Jack H. Grant [a,f], Sophie Harte [a,e], Brad T. Hudson [a,c], Maria Sablik [a,b], Nicholas A. Badcock [g,h] and Dorothy V.M. Bishop [a,*]

[a] Department of Experimental Psychology, University of Oxford, UK
[b] School of Human and Behavioural Sciences, Bangor University, UK
[c] Department of Psychology, Lancaster University, UK
[d] Department of Psychology, University of Essex, UK
[e] Deafness, Cognition and Language Research Centre, University College London, UK
[f] Lincoln Medical School, University of Lincoln, UK
[g] School of Psychology, University of Lincoln, UK
[h] School of Psychological Sciences, University of Western Australia, Australia

ARTICLE INFO

Article history:
Protocol received 05 September 2020
Protocol approved 17 December 2020
Received 2 April 2022
Reviewed 25 April 2022
Revised 24 May 2022
Accepted 27 May 2022
Action editor Chris Chambers
Published online xxx

ABSTRACT

Background: Most people have strong left-brain lateralisation for language, with a minority showing right- or bilateral language representation. On some receptive language tasks, however, lateralisation appears to be reduced or absent. This contrasting pattern raises the question of whether and how language laterality may fractionate within individuals. Building on our prior work, we postulated (a) that there can be dissociations in lateralisation of different components of language, and (b) these would be more common in left-handers. A subsidiary hypothesis was that laterality indices will cluster according to two underlying factors corresponding to whether they involve generation of words or sentences, versus receptive language.

Methods: We tested these predictions in two stages: At Step 1 an online laterality battery (Dichotic listening, Rhyme Decision and Word Comprehension) was given to 621 individuals

Abbreviations: CBFV, cerebral blood flow velocity; fTCD, functional transcranial Doppler ultrasound; LIz score, laterality index expressed as individualised z-score.

* *Corresponding author.*
E-mail address: dorothy.bishop@psy.ox.ac.uk (D.V.M. Bishop).
[1] Joint first author.
[2] Current address: Department of Experimental Psychology, University College London, 26 Bedford Way, London WC1H 0AP, UK.

https://doi.org/10.1016/j.cortex.2022.05.013

Keywords:
Language laterality
Handedness
Dichotic listening
Visual half-field
Functional transcranial Doppler ultrasound

(56% left-handers); At Step 2, functional transcranial Doppler ultrasound (fTCD) was used with 230 of these individuals (51% left-handers). 108 left-handers and 101 right-handers gave useable data on a battery of three language generation and three receptive language tasks.

Results: Neither the online nor fTCD measures supported the notion of a single language laterality factor. In general, for both online and fTCD measures, tests of language generation were left-lateralised. In contrast, the receptive tasks were at best weakly left-lateralised or, in the case of Word Comprehension, slightly right-lateralised. The online measures were only weakly correlated, if at all, with fTCD measures. Most of the fTCD measures had split-half reliabilities of at least .7, and showed a distinctive pattern of intercorrelation, supporting a modified two-factor model in which Phonological Decision (generation) and Sentence Decision (reception) loaded on both factors. The same factor structure fitted data from left- and right-handers, but mean scores on the two factors were lower (less left-lateralised) in left-handers.

Conclusions: There are at least two factors influencing language lateralization in individuals, but they do not correspond neatly to language generation and comprehension. Future fMRI studies could help clarify how far they reflect activity in specific brain regions.

1. Introduction[3]

Cerebral lateralisation for language has been studied both in populations and within individuals. At the population level, it is well-established that, for most people, language generation is predominantly controlled by the left hemisphere of the brain. There is individual variation, and a minority of people have right hemisphere language or do not show clear bias to one side. Numerous sources of evidence converge to show that atypical language laterality is more common in left-handers (around 70% left-lateralised) than right-handers (around 95% left-lateralised; Carey & Johnstone, 2014; Knecht, Deppe, Dräger, Bobe, Lohmann et al., 2000; Rasmussen & Milner, 1975, 1977; Vingerhoets, 2019). Strong left-lateralisation is not seen for all aspects of language, however. As will be reviewed below, functional brain imaging has shown that on some language tasks left-lateralisation at the population level is either reduced or absent. If we regard language lateralisation as a single dimension it may be tempting to conclude that a language task that is not strongly lateralised at the population level is likely to be a noisy or invalid measure (cf. Bethmann, Tempelmann, De Bleser, Scheich, & Brechmann, 2007; Sørensen & Westerhausen, 2020). In practice, little attention has been given to individual differences in language functions that are not strongly lateralised. Vingerhoets (2019) noted that to date few studies have distinguished between left-lateralised, right-lateralised and bilateral phenotypes within a single lateralised function, and few have compared laterality across different functions.

Tasks that do not show a population bias to left or right may nevertheless be lateralised within individuals – but with absent or reduced population bias to left-sided functioning. For example, if we have a language task where there is a 50:50 mixture of left- and right-lateralised individuals, the population mean will indicate no lateralisation, suggesting that both hemispheres participate in the task in individuals. However, a subset of individuals may nevertheless be reliably lateralised to one side or the other, but with equal proportions being left-lateralised or right-lateralised. In that case, we should find people who have different language functions mediated by opposite hemispheres. This notion is compatible with suggestive evidence that some people have discrepant language lateralisation for different tasks measured using fMRI (e.g., Lee, Swanson, Sabsevitz, Hammeke, Scott, Possing, & Binder, 2008; Ramsey, Sommer, Rutten, & Kahn, 2001) or for Q2 different brain regions when performing a single task (Bethmann et al., 2007). It has, however, been difficult to draw firm conclusions due to a lack of reliable and validated laterality measures and well-powered studies in this area. An additional problem is the lack of available data on individual participants in the majority of studies that have used fMRI, electrophysiology or behavioural methods to assess laterality.

It is important to clarify the nature of atypical lateralisation, because cerebral asymmetry has clinical implications for such issues as epilepsy surgery and recovery from aphasia, as well as informing our understanding of brain-behaviour associations, and the neurobiological basis of the human language capacity. For these reasons we designed a multi-centre study to quantify several measures of cerebral asymmetry using behavioural and physiological measures, in a sample enriched with left-handed participants. We acquired a large dataset of reliable language laterality measures, including behavioural methods and functional transcranial Doppler ultrasound (fTCD), which is shared following the principles of open research.

[3] In our Stage 1 Report (registered at https://osf.io/p8k2b) we gave the background for our preregistered hypotheses and analysis plan for this study. In the current paper, with the exception of the Abstract, material prior to Participants has the same wording as the Stage 1 Report, except for minor typographical changes (e.g., verb tense), plus some sentences where departures to preregistration are made explicit. The latter are in italic font. Some preregistered material has been moved after the Results section to improve logical flow; this is shown in bold italic.

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

1.1. Individual differences in language lateralisation

Woodhead, Bradshaw, Wilson, Thompson, and Bishop (2019) and Woodhead, Thompson, Karlsson, and Bishop (2021) compared language lateralisation for a range of tasks designed to engage different aspects of language functioning. These studies used fTCD, which quantifies lateralisation by directly comparing blood flow in left and right middle cerebral arteries during performance of an activation task. A group of 31 left-handers and 43 right-handers performed six language tasks twice on two occasions. Tasks were selected based on the dual-stream model of Hickok and Poeppel (2007), which postulates that language functioning engages two parallel processing streams. The dorsal pathway feeds forward to the inferior frontal gyrus where phonological representations are translated into articulation and is strongly left-lateralised. The ventral pathway, involving the middle and inferior temporal gyri, maps phonological representations onto lexical conceptual representations, and is thought to have weak or absent left-sided dominance. The two streams are not independent; they are both connected to a left-lateralised combinatorial network. Woodhead et al. (2019, 2021) selected tasks involving List Generation (covertly reciting overlearned sequences such as days of the week, months of the year) and Phonological Decision (judging if pictured words rhymed) to index dorsal stream activity. To engage the ventral pathway they selected tasks involving Semantic Decision (judging if pictured words belonged in the same semantic category) and Syntactic Judgement (judging whether a series of nonsense words such as "The tarben yipped a lev near the kruss" had grammatical structure). Tests of Sentence Generation (covertly describing a picture) and Sentence Decision (selecting a picture to match the meaning of a spoken sentence) were predicted to involve both pathways. Woodhead et al. (2019, 2021) found that a bifactor model did better than a single factor model at accounting for covariances between laterality indices, but the two factors did not divide neatly according to ventral/dorsal stream predictions. Contrary to prediction, the List Generation task did not show high loadings on either factor, despite taxing articulatory processes, though it should be noted this task also had much lower test-retest reliability than other tasks ($r_s = .33$). The strongest laterality was seen for the Sentence Generation task, which indexed the first factor. This factor had significant loadings from Phonological Decision, Semantic Decision and Sentence Decision. All of these tasks were left-lateralised overall, though to varying extents. The second factor had loadings from Sentence Decision (which was left-lateralised) and Syntactic Decision (which was not lateralised). The two factors were highly correlated in right-handers, but in left-handers they were less well correlated. The tentative interpretation of these results was that in some individuals (primarily left-handers) there can be a dissociation between laterality for language generation (factor 1) and comprehension (factor 2). The pattern of results also suggested that word retrieval may be the key process characterising factor 1, rather than the articulatory aspect of speech production. Accordingly, we refer to this as 'language generation' rather than 'production'. It was noteworthy that the tasks loading strongly on factor 2 were the only two tasks that involved auditory presentation of stimuli.

An important feature of the data was that test-retest reliability of the laterality index was as high for tasks that were poorly lateralised as for those that were strongly lateralised (Woodhead et al., 2019, 2021). This challenges the interpretation that these are tasks where both hemispheres are equally involved. If that were the case, the true laterality index would be zero for all people, and any individual variation would just be noise, so test-retest reliability would be low. Instead, there were stable individual differences for "bilateral" tasks, but with equal probability of bias to left or right. This does not preclude the possibility that some people have true "bilateral language", i.e., equal involvement of both hemispheres, but it does challenge the idea that tasks that show no bias at the population level invariably mean there is no bias in individuals. Further evidence comes from a fTCD study by Woodhead, Rutherford, and Bishop (2018), which included a List Generation task that was not significantly lateralised. The LI from this task was nevertheless significantly correlated with laterality indices for word and sentence generation, both of which showed the usual left-hemisphere bias. Overall, these observations suggest that there are meaningful, stable, individual differences in degree of lateralisation, even for tasks that show no bias at the population level; this is consistent with the notion that different language functions may be primarily mediated by opposite hemispheres in some individuals (Bishop, 2013).

The primary goal of the current study is to consolidate the findings of Woodhead et al. (2019, 2021) by replicating and extending the findings of dissociated language functions, using some new tasks. We studied a large sample using online behavioural laterality assessment, plus a smaller subset of these individuals assessed with fTCD.

1.2. Methodological considerations

Most contemporary studies of brain lateralisation use fMRI, which provides information about localisation as well as laterality of brain activation. Estimates of lateralisation from fMRI are dependent on the experimental task, the specific brain region, and the selection of a baseline task. Decisions about how to quantify laterality, and selection of statistical thresholds can lead to different estimates of group and individual asymmetry (Bradshaw et al., 2017; Seghier, 2008; Wilke & Lidzba, 2007). In addition, fMRI has poor temporal resolution and is expensive enough to preclude routine studies using large numbers of participants. In contrast, fTCD—the method used by Woodhead et al. (2019, 2021)—allows for direct comparison of blood flow in the left and right middle cerebral arteries, and has good temporal resolution. Although fTCD cannot provide information about which brain regions are active, it is considerably less expensive than fMRI and is portable. FTCD has also been validated using the gold standard Wada test of language lateralisation (Knecht, Deppe, Ebner, Henningsen, Huber et al., 1998). A third approach to the study of functional asymmetry, which dates back to the 1960s, involves inferring which hemisphere is more engaged in behavioural tasks when visual or auditory stimuli are presented in a way that preferentially engages one hemisphere (reviewed in Bryden, 1982). Accuracy or response time

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

measures can be used to provide an indication of left- or right-sided bias, both at the group level and in individuals. We have recently shown that good quality data can be obtained with some behavioural tasks using online administration, which makes it feasible to assess very large samples (Parker, Woodhead, Thompson, & Bishop, 2021).

In principal, it would be of value to use all three approaches with the same set of participants, to obtain convergent evidence from methods that make different assumptions and use different approaches to assess laterality. Ultimately, we aim to adopt that approach: here we made a start on that goal with a study of individual differences in language lateralisation that uses the last two of these methods: behavioural testing in a large sample, followed by fTCD with a subset of the same participants. We first describe the rationale for selecting specific measures: the tasks are described in greater detail under Methods.

1.3. *Behavioural measures*

Our selection of behavioural measures was guided by methodological and theoretical considerations. In terms of methodology, we have been exploring the use of online administration for behavioural laterality tasks (Parker et al., 2021). Given our interest in the nature of bilateral language, we do not regard it as important that a task shows a population bias in lateralisation, provided that test reliability is strong, indicating stable individual differences (see Positive Controls, below). From a theoretical perspective, we have a particular interest in contrasting tasks that involve language generation versus receptive language, while noting that this distinction is not always clearcut, as many receptive tasks can involve covert language generation.

One of the first behavioural tasks used to study language lateralisation is dichotic listening, where a person hears simultaneous streams of words or speech sounds in left and right ears. Under binaural presentation, the contralateral auditory pathways take precedence and the ipsilateral pathways are suppressed. Thus sounds presented to the right ear have a more direct access to the left hemisphere speech systems than those played to the left ear, with a strong bias to report items from the right ear being present at a population-level. Good reliability (above r = .75) was found for a dichotic listening task administered via a mobile app (Bless et al., 2013) and we found similarly high levels of reliability for this task using online presentation (Parker et al., 2021). Dichotic listening laterality is not, however, strongly predictive of language laterality as measured by fMRI (Bethmann et al., 2007; although see Sørensen & Westerhausen, 2020, for a reappraisal). This lack of specificity could mean that factors such as attentional bias affect performance, invalidating the test as a measure of language laterality, particularly in individual participants. However, another possibility is that dichotic listening is a good measure of lateralisation of receptive language, but it may be dissociable from laterality for language generation.

Another type of behavioural method to assess lateralisation involves visual presentation. Stimuli are briefly placed in the left or right visual half-fields, which project primarily to the contralateral hemisphere. This method has long been used to assess language laterality, either using written words or pictures as stimuli (Bryden, 1982). Laterality indices show a right visual field advantage (VFA) at the population level, but results depend crucially on specific aspects of task design. Laterality indices from such tasks do not, however, necessarily correlate highly with dichotic listening (Voyer, 1998). Hunter and Brysbaert (2008) argued that one needs a visual half-field task involving speech production to obtain good prediction of language laterality as measured by fMRI. This hypothesis fits with our theoretical perspective: a visual half-field task that involves language generation would be expected to be better than dichotic listening for predicting lateralisation of word generation as measured by fTCD. Van der Haegen and Brysbaert (2018) reported reliability for three visual laterality tasks in a sample of 50 left-handers tested on two occasions, with test-retest correlations ranging from .49 (optimal viewing position - OVP- for written words), .77 (visual half-field with pictures) to .83 (visual half-field with words). Parker et al. (2021) developed a new Rhyme Decision task that involved covert naming, but reliability was below a pre-specified cutoff for acceptability of .65 (Spearman $r_s = .63$). We subsequently gathered pilot data on a modified version of the task for 15 left-handers and 15 right-handers, and obtained split-half reliability of .74. Contrary to our expectation, task performance was not significantly lateralised in either left-handers or right-handers, but the good reliability indicates it measures a stable individual difference.

The pilot study also gathered data on two further tasks designed to tap into more receptive aspects of language: the OVP task, and a new Word Comprehension task, that simply involved selecting which of two laterally-presented pictures matched a spoken word. Reliability of the OVP task was relatively poor, but the laterality index from the Word Comprehension task had split-half reliability of .74, again suggesting there are stable individual differences in lateralisation. In our pilot data, the Word Comprehension task was significantly lateralised, but in the opposite direction to prediction, i.e., with better performance for pictures shown in the left compared to the right visual field. Given our goal of using reliable tasks that involve language generation or receptive language, regardless of lateral bias shown on the tasks, we decided to focus on dichotic listening, Rhyme Decision and Word Comprehension (see Methods). In practice, our chosen language tasks differ in ways other than the generative/receptive distinction: one visual half-field task, for instance, involves written rather than spoken language, and some tasks require accessing meaning whereas others do not. It would not be possible to design a battery that completely controlled for all task variables; rather we planned to administer tasks with diverse characteristics, predicting that generation versus reception of language will determine which laterality indices form a common factor.

1.4. *Functional transcranial Doppler ultrasound tasks*

The fTCD tasks included identical or closely similar versions of four of the tasks previously used by Woodhead et al. (2019, 2021), all of which had good test-retest reliability. In that study, two tasks (Sentence Generation and Phonological Decision) loaded primarily on factor 1 (language generation), and

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

two tasks (Sentence Decision and Syntactic Decision) loaded primarily on factor 2 (receptive language). For each factor, one additional task was used: the gold standard Word Generation task with letter stimuli for language production (Knecht et al., 1998), and a new Word Decision task for receptive language. The latter task used the same materials as the online Word Comprehension task described above; although the two tasks were closely similar, we gave them different names to make it easier to distinguish the behavioural and fTCD versions.

1.5. Positive Controls

If there are no significant correlations between different tasks it is important to demonstrate that this is not simply due to use of inadequate measures or impact of uncontrolled unwanted variables. Demonstration of good reliability of laterality indices in effect provides a positive control. Parker et al. (2021) found very weak intercorrelations between a set of online tasks, despite test-retest reliability for individual tasks of .7 or above. Woodhead et al. (2019, 2021) showed that for the six tasks used in their fTCD study, dissociations between LIs for different tasks could not be attributed to weak reliability, as all laterality indices (LIs) except list generation showed test-retest reliability (rs) of .6 or more). In the current study, we did not have resources to test all participants twice, but planned to repeat the online tests for a subset of 50 individuals (50% left-handers). In addition, split-half reliability using alternate items was assessed for all measures.

1.6. Sampling approach

One reason for uncertainty about the phenomenon of dissociated language functions is that laterality measures follow a strongly skewed distribution, and people with dissociated or atypical lateralisation are, by definition, rare (Johnstone et al., 2020; Mazoyer et al., 2014). Some researchers with an interest in atypical lateralisation have focussed exclusively on left-handers, which gives a higher yield of such individuals (Gerrits, Verhelst, & Vingerhoets, 2020; Van der Haegen & Brysbaert, 2018). Given our findings that handedness may influence patterns of association and dissociation of lateralised language functions, we planned to recruit both left- and right-handers in a 2:1 ratio, to give adequate power to detect such differences.

A further complication is that left-handers do not form a uniform group. Over many years, various suggestions have been made about possible subdivisions between types of left-handers: in particular it has been proposed that right-sided language lateralisation is associated with extreme left-handedness (Knecht et al., 2000; Mazoyer et al., 2014). Another common idea is that familial left-handedness distinguishes between subtypes of left-handers, or may identify a genetic predisposition to left-handedness in right-handers (McKeever & Vandeventer, 1977). This notion remains popular, although empirical support is weak (Orsini, Satz, Soper, & Light, 1985). Indeed, it has been criticised for making no sense in relation to genetic models of handedness (Bishop, 1980; 1990), which attribute a relatively minor causal role to genes and a high contribution from chance factors. We did not make strong predictions about variation within left-handers, but we gathered data on strength and familiality of handedness that will allow for exploratory analyses of this topic.

We describe below the rationale for sample size determination. With online testing, we gathered a large number of participants, which is the basis for a preliminary test of the 'dissociable language laterality' hypothesis. The initial sample was recruited according to handedness, with the goal of having 300 left-handers and 150 right-handers.

In the second phase of the study, we compared findings from the online measures with those obtained using direct measures of brain lateralisation from fTCD on a subset of the initial sample. We aimed to test around half the sample on fTCD as well as online methods, as simulations indicated that a sample with 112 left-handers and 112 right-handers would be adequately powered to test our hypotheses (see Sampling and Analysis plan, below). Note that online test results were not used to select individuals for phase 2: the aim was to test all available participants until our quota of left- and right-handers was met.

1.7. Research questions

The overarching question is whether there are cross-hemispheric dissociations in lateralisation of different language functions, and if so whether there are separable dimensions of laterality for tasks that primarily implicate language generation and receptive language. A positive answer to this question would challenge the conventional conceptualisation of language lateralisation as a unitary dimension, and support instead the dissociable language laterality hypothesis.

A subsidiary question is whether dissociation between laterality dimensions is more characteristic of left- than right-handers.

A final question is whether online behavioural measures are comparable to direct measures of cerebral blood flow in indexing language laterality. It is generally assumed that both types of laterality measurement are indexing the same underlying bias, but the nature of what is measured is very different: facilitation of processing material on one side for behavioural measures, and lateralised increase in blood flow through the middle cerebral artery in the other.

1.8. Hypotheses and predictions

1.8.1. Online behavioural measures

1. It is predicted that the pattern of correlation between laterality indices from online measures will reflect the extent to which they involve language generation, rather than whether they involve spoken or written language. Thus we anticipated dissociation between the Rhyme Decision task, which requires covert speech production, and the Word Comprehension task and Dichotic Listening tasks, which do not. We further anticipated that dissociations between tasks are not accountable for in terms of low reliability of measures – i.e., correlations of laterality indices between tasks will be lower than split-half reliability of the measures.

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

1.8.2. FTCD measures

2. The same hypothesis predicts that the fTCD data will fit a model where 'language generation' tasks cluster together on one factor, and 'receptive' language tasks on a second factor. The factors will be correlated, but the fit of a two-factor model will be superior to a single-factor model.
3. From our hypothesis that handedness affects language laterality, following Woodhead et al. (2021), we predicted that better model fit will be obtained when different parameters are estimated for left- versus right-handers, compared with when all parameters are equated for the two handedness groups.
4. The same hypothesis leads to the further prediction that on categorical analysis, individuals who depart from left-brained laterality on one or more tasks will be more likely to be left-handed than those who are consistently left-lateralised.

1.8.3. Relationship between fTCD and behavioural laterality indices

5. Our predictions depend on online and fTCD measures indexing the same lateralisation processes. On this basis we predict that the laterality profile obtained with the online language battery will be significantly associated with the profile seen with the direct measurement of cerebral blood flow using fTCD, with laterality on dichotic listening and Word Comprehension relating more strongly to receptive language tasks, and Rhyme Decision to language generation tasks.

2. Methods

2.1. Criteria for participants

Our original stage 1 flowchart for participant recruitment is now presented as Fig. 3 below, showing both the original planned sample size and the obtained sample size.

The inclusion criteria were as follows:

- Aged 16–50 years. The younger age limit avoids developmental change in language skills affecting performance, and the upper limit makes it less likely that bone density will make it difficult to find a Doppler signal.
- Normal or corrected-to-normal vision and normal hearing
- Native English proficiency
- Access to a laptop or desktop computer with stereo headphones for use in the online testing. N.B. it is not possible to do the online tests on a tablet or phone.

The exclusion criteria were as follows:

- A history of psychiatric or neurological illness
- A history of developmental language disorder, dyslexia or autism.
- A history of dyspraxia
- Unwillingness to travel to one of the testing sites for Step 2 of the study
- Contraindications or unwillingness to participate in fMRI in future parts of the study.

The initial screening questionnaire was administered on the Gorilla platform (www.gorilla.sc; Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020), and is provided in Supplementary Material 1 (Available at: https://osf.io/g9tqh/). Note that we did not exclude those who speak more than one language, provided they met our stringent criteria for native-level competence in English (see below). We also gathered information on bilingualism/multilingualism in the initial demographic questionnaire, so it would be possible to determine if this had any impact on results. Participants were told that this test was part of a multistage study and they might be invited back for in-person testing. Although there was no obligation on them to do so, if they were in principle willing, they had the opportunity to provide contact details.

Although this study did not include fMRI, we prioritised participants who were likely to be eligible for fMRI in a future phase of the research. The initial screening questionnaire checked whether participants were in principle willing to return for an MRI scan, and whether they were aware of any contraindications to being scanned.

If the participant passed the screening questionnaire, they were invited to complete an online consent form prior to starting the online testing session. Those who took part in the second session (in-person testing with fTCD) completed an additional written consent form for that session.

2.2. Procedure

There were two stages to the project. Participants who proceeded beyond the initial screening were invited to complete the online testing (Step 1), and a subset of participants were invited back for the fTCD session (Step 2). Participants received course credit or were paid in accordance with guidelines at their local testing centre (at least £8 per hour).

2.3. Step 1: Online testing

After passing the screening questions and completing an online consent form, participants continued with the online testing via the Gorilla platform (www.gorilla.sc; Anwyl-Irvin et al., 2020). In this session participants completed the following (described in more detail below):

- Demographics questionnaire
- Edinburgh Handedness Inventory (Oldfield, 1971)
- A test of ocular dominance and footedness
- Measures of language proficiency.
- Tests of language laterality
  - Rhyme Decision
  - Word Comprehension
  - Dichotic Listening

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

Many of these tests have been reported previously, and have been made available for reuse: https://gorilla.sc/openmaterials/104636.

2.3.1. *Demographics questionnaire*

The demographics questionnaire is shown in Supplementary Materials 2 (Available at: https://osf.io/g9tqh/). Participants were asked to report age, gender, years in education, and whether they were bilingual. They were also asked about their own hand and foot preference, and left-handedness in first degree relatives (parents and siblings – see question 7). The latter information was used to compute a proportional familial sinistrality index (Corey & Foundas, 2010).

2.3.2. *Edinburgh Handedness Inventory (EHI)*

Participants completed the Edinburgh Handedness Inventory (Supplementary Material 3 (Available at: https://osf.io/g9tqh/); Oldfield, 1971) in order to quantify handedness on a continuum. For 10 activities, participants indicated their hand preference on a 5-point scale (right hand strongly preferred, right hand preferred, no preference, left hand preferred, left hand strongly preferred). This questionnaire was selected for compatibility with prior studies relating language laterality to hand preference.

2.3.3. *Test of ocular dominance*

A version of the Porta test (Porac & Coren, 1976) that is suitable for online testing was administered to determine each participant's eye dominance in central gaze (i.e., when looking straight ahead). This test classifies participants as being either left or right eye dominant.

2.3.4. *LexTALE*

The Lexical Test for Advanced Learners of English (LexTALE; Lemhöfer & Broersma, 2012) was used to assess level of English vocabulary knowledge. Participants judge the lexical status of 60 letter strings (word or non-word). Forty are real English words and 20 are non-words. To correct for the unequal proportion of words and non-words, LexTALE scores are calculated as [((number of words correct/40 × 100) + (number of nonwords correct/20 × 100))/2]. Following the norms provided by Lemhöfer & Broersma, those scoring below 80 are not eligible for inclusion in the online testing or fTCD. *N.B. In practice, we did not implement this exclusion, for reasons stated below.*

2.3.5. *Games with words test*

The Games With Words test (Hartshorne, Tenenbaum, & Pinker, 2018) was used to screen participants for adequate understanding of English grammar. The first 8 items involve participants reading a sentence, such as "The dog was chased by the cat", and deciding which of two pictures presented below the text matches the sentence. The two pictures in this example include a dog chasing a cat and a cat chasing a dog. Items 9–35 were four-alternative forced choice questions where participants select which of four sentences sounds most natural: e.g., (1) "What age are you?", (2) "How age are you?", (3) "How old are you?", and (d) "What old are you?". To be included as having native English speaker proficiency, participants need to make no more than 3 errors on this test: Hartshorne et al. (2018) reported that most monolingual English-speakers performed close to ceiling and very few made more than 3 errors. *N.B. In practice, we did not implement this exclusion, for reasons stated below.*

2.3.6. *Rhyme decision*

A modified version of a Rhyme Decision visual half-field task reported by Parker et al. (2021) was administered to determine brain lateralisation for language generation. This involved participants judging which of two parafoveal images rhymed with a foveally presented word. The laterality index from the original task had test-retest reliability of r = .63, and the overall lateralisation effect, though significant, was small. We modified the task from the original with the aim of improving its psychometric properties: first, we removed trials in which neither of the pictures rhymed with the target word, as these were potentially confusing. Second, we increased the distance between the centrally presented word and parafoveal images, to ensure the image is projected exclusively, at least initially, to the contralateral hemisphere. Pilot testing with 30 participants obtained split-half reliability of r = .74 with this version. *Note, however, lower reliability was found with our main sample, as reported in Results.*

2.3.6.1. Materials. Written stimuli consisted of twenty-six monosyllabic written words (e.g., bite). Stimulus pairs were created so that each written word was paired with an image with a name that rhymed (e.g., kite). For each of the 26 word–image pairs there was a corresponding pair whose words did not rhyme with the first pair (e.g., the corresponding pair for bite-kite was more-door), see Supplementary Materials 4 (Available at: https://osf.io/g9tqh/). On each trial, stimuli were presented such that the written word was accompanied by one rhyming image and one non-rhyming image (e.g., kite-bite-door). All possible combinations were included such that each pairing constituted four individual items. Thus, there were 52 unique stimuli, and all images appeared both as rhymes and non-rhymes (see Parker et al., 2021, for further detail).

Word stimuli were presented in 28 pt. black Courier New font on a white background. The images were displaced at 7.9 degrees of visual angle from the point of fixation. Gorilla's screen scaling tool was used to maintain consistency of stimulus size across browsers and computers.

2.3.6.2. Procedure. Participants completed a familiarisation procedure where they viewed each image pair. The images were shown with their name presented in text below the image to ensure that participants used the appropriate word when making a rhyme decision. They then completed 208 trials of the Rhyme Decision task (four blocks of 52 stimuli). Each trial began with a central fixation cross which was visible for 800 msec after which a foveally presented word appeared for 200 msec. Two bilateral images then appeared for 150 msec. At stimulus offset, participants indicated whether the centrally presented word rhymed with the image present to the left or right visual field by pressing S for the left visual field and K for the right visual field. Participants' responses triggered the next trial. Accuracy and response times were recorded.

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

2.3.7. *Word comprehension task*

A novel online Word Comprehension task was administered to determine lateralisation of receptive language. This task involved indicating which of two semantically related parafoveal images matched an orally presented word stimulus. Pilot testing with 30 participants and one block of stimuli obtained split-half reliability of r = .76. *Note, however, lower reliability was found with our main sample, as reported in Results.*

2.3.7.1. Materials. A total of 108 experimental images were selected from the MultiPic databank (Duñabeitia et al., 2018). As Duñabeitia et al. had participants name stimuli, we were able to select images where at least 50% of English participants generated the intended name; M = 90.1%, SD = 13.06. The names of the images were high frequency according to Zipf scores from the SUBTLEX-UK database (Van Heuven, Mandera, Keuleers, & Brysbaert, 2014); M = 4.4, SD = .41, minimum = 3.8. Each image was paired with another to form a semantically related pair, amounting to 54 pairs (see Supplementary Materials 5 (Available at: https://osf.io/g9tqh/). Latent semantic analysis ratings were acquired using the LSA CU Boulder web-interface (http://lsa.colorado.edu/). LSA scores ranged from .26 to .87; M = .46, SD = .87. We aimed to avoid errors arising from visual confusion between pictures, and to this end, nine raters rated the pairs of images for visual similarity on a 5 point scale (1: very similar to 5: not all similar). Generally, image pairs were rated as being visually distinct; M = 4.13, SD = .47. Audio files of the spoken name for each picture were created using Google cloud text-to-speech (https://cloud.google.com/text-to-speech), using a male, British voice.

Each image pair was presented a total of four times, with each image twice in either the left or right visual field. The name of each image was presented twice: once when the image was in the right visual field, and once when the image was in the left. See Fig. 1 for an example. The images were displaced at 7.9 degrees of visual angle from the point of fixation. Gorilla's screen scaling tool was used to maintain consistency of stimulus size across browsers and computers. In total, 208 trials were presented across four blocks.

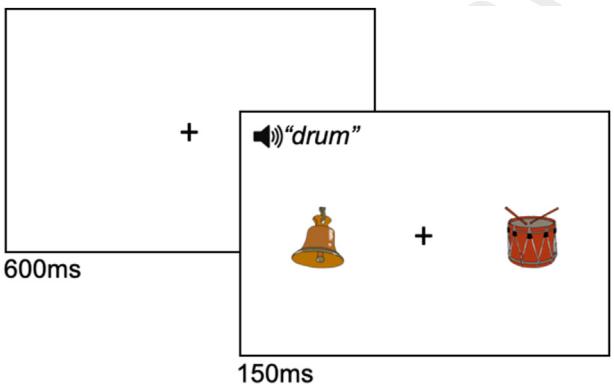

Fig. 1 – Schematic illustration of a trial on the Word Comprehension task. First, participants are presented with a fixation cross for 600 msec. A word, in this case 'drum', is presented aurally, along with two semantically related parafoveal images that are visible for 150 msec. Participants then indicate which of the two images matched the word. In this case participants would indicate that the word matched the image on the right. This figure has been modified from the preregistered version for clarity.

2.3.7.2. Procedure. Participants completed a familiarisation procedure (where each picture was shown along with its spoken name) and a number of practice trials. They then completed 208 experimental trials (four blocks of 52 stimuli). Each trial began with a central fixation cross which was visible for 600 msec. The target word was then presented aurally along with two semantically related parafoveal images that were visible for 150 msec. Participants indicated whether the oral word matched the image in the left or right visual field by pressing Q for the left visual field and P for the right visual field. Participants' responses triggered the next trial. Accuracy and response times were recorded.

2.3.8. *Dichotic listening task*

An online Dichotic Listening task was administered to assess the lateralisation of speech perception (Hugdahl & Andersson, 1986). On each trial, participants heard two consonant-vowel (CV) auditory stimuli simultaneously to each ear. The stimuli have previously been administered online via an app (Bless et al., 2013) and Gorilla (Parker et al., 2021) and the task has good test-retest reliability (r = .78).

2.3.8.1. Materials. Six stop-consonants (/b/, /d/, /g/, /p/, /t/, /k/) were combined with the long vowel /a/ to create consonant-vowel stimuli. Stimuli are paired in all possible combinations and played in each sound channel (e.g., /pa/-/ga/). This resulted in 36 unique pairings (including pairs with the same sound repeated).

2.3.8.2. Procedure. To ensure adequate headphone use, participants were screened on two measures. The first, described by Woods, Siegel, Traer, and McDermott (2017), involves participants deciding which of three pure tones is the quietest. One of the three tones is played 180° out of phase, so this task is difficult to perform through speakers but relatively easy with headphones. The second task was a stereo check developed by Parker et al. (2021). This task involves participants listening to a sound played in a single channel and reporting whether the sound was played to the left or the right ear via a button press. Each task had six trials. We excluded participants who scored less than 4 correct on each.

Participants completed three blocks of the Dichotic Listening task. This amounted to 96 trials when excluding homonyms. Our decision to use three blocks was based on the previous observation that there is not much improvement in reliability after 85 trials when using the Dichotic Listening task (Parker et al., 2021). Each trial began with a fixation cross, which was presented for 250 msec. Participants then heard the sound pairs and reported the syllable that they heard. If they heard two different syllables, participants were instructed to report the sound that they heard the most clearly. Responses were made by clicking a button which corresponded to one syllable. The response triggered the start of the next trial. The side of selected response and response times was recorded. Errors can occur on this task if the participant

selects a syllable that was not presented to either ear: these were rare, but were recorded and were used to exclude those who performed poorly.

2.3.9. *General procedure*

The study implemented a within-subjects design where participants completed all online tasks. When completing the battery, participants were instructed to sit approximately 50 cm from the screen. Participants completed four blocks of the Rhyme Decision (A) and Word Comprehension (B) tasks (52 trials each). They completed three blocks of the Dichotic Listening (C; 36 trials each). Blocks of different tasks were interspersed in a quasi-random order (i.e., ABC, BAC, CAB). Other online tasks relating to different studies were interspersed to avoid boredom and maximise efficiency of data collection.

2.4. *Step 2: Functional transcranial Doppler ultrasound (fTCD)*

A subset of participants were given six tasks using fTCD that could be administered in a single session of about 90–120 min. These participants were selected on the basis of handedness and willingness and ability to travel for in-person testing to one of the *six test centres (Oxford, University College London, Bangor, Lincoln, Lancaster, University of Western Australia)*.

The tasks were designed to cover a range of language functions in as standard a format as possible. Four of the language tasks (tasks B, C, E and F below) were based on tasks used by Woodhead et al. (2019). We omitted the List Generation task used by Woodhead et al. (2019), as it showed poor test-retest reliability. Instead, we used a shortened version of the gold standard Word Generation task (Knecht et al., 2000). As detailed below, the basic procedure for this task was the same as used by Knecht et al., but with fewer trials and a shorter rest period between trials. The sixth task was a new Word Decision task, selected to act as a third indicator of receptive language.

The six tasks are described below.

2.4.1. *Task A: Word generation*

On each trial, the task was to silently generate words that begin with a specified letter, and subsequently report them when a cue is given. The task used 18 letters that commonly begin English words (S, C, P, D, T, B, R, A, E, F, G, H, I, L, M, U, O, W).

2.4.2. *Task B: Sentence generation*

This task was based on Mazoyer et al. (2014). Participants were shown a line drawing in each trial and asked to produce a meaningful sentence to describe it, following a prescribed simple structure, (e.g., "The boy in the hat flew the kite"). The 18 stimuli were selected from those used by Woodhead et al. (2018).

2.4.3. *Task C: Phonological Decision*

Following a familiarisation task, for each trial, participants decided whether the names of pairs of pictures rhymed. The design of the original task from Woodhead et al. (2019) was supplemented by adding a number of black and white drawings from the MultiPic database (Duñabeitia et al., 2018), in order to create stimuli for three new trials.

2.4.4. *Task D: Word Decision*

Participants were presented with pairs of pictured items on each trial, and were asked to press a key to indicate which one matched a spoken word. The pictures were from the same semantic category. The picture pairs and audio stimuli for this task were the same as those used in the online Word Comprehension task (see below for further details).

2.4.5. *Task E: Sentence Decision*

Two pictures were presented on each trial, and the task was to determine which one matched a spoken sentence. Items were based on picture pairs taken from the Test for Reception of Grammar – 2 (Bishop, 2003), using distractors that differed in syntactic arrangement of words. As we used slightly more trials than Woodhead et al. (2019), additional pictures and sentences were devised in the same way. New spoken sentences were created for all items using Google text to speech, with a male British voice.

2.4.6. *Task F: Syntactic Decision*

On each trial, participants were presented with a sequence of words and non-words, and asked to judge if they formed a plausible "Jabberwocky" sentence with correct syntactic structure. Simultaneous spoken and written presentation was used. The stimuli used by Woodhead et al. (2019) were supplemented with additional Jabberwocky stimuli from Fedorenko, Hsieh, Nieto-Castañón, Whitfield-Gabrieli, and Kanwisher (2010).

Timings for the tasks are shown in Fig. 2. All stimulus materials for the tasks are available on OSF: (https://osf.io/g3qms/?view_only=a6c36957ffba4bc39232d9265ea13dd8). We previously used 15 trials of each task but increased this to 18 trials for the current study, to ensure that there were sufficient stimuli for reliable estimation of a laterality index, even if some trials were lost due to recording problems. Previously we presented tasks with an inter-stimulus interval (ISI) of 33 sec (including 20 sec of the task and 10 sec of rest). We increased this by extending the rest period to 15 sec. Hence, the ISI was 38 sec and each task lasted 11 min 24 sec in total. The overall testing time (excluding set-up, practice trials and breaks) was 68 min 24 sec.

Task order was counterbalanced between participants to avoid order effects using a replicated Latin square design using a customised script in R Studio. Details of task A are described by Woodhead et al. (2018), and Tasks B, C, E and F by Woodhead et al. (2019). Task D was designed for this study and is described below.

Task D (Word Decision) used the same 54 picture pairs and audio stimuli that were used for the online Word Comprehension task. The procedure for this task in fTCD has been designed to match that of the Sentence Decision task (task E). In each epoch, a pair of drawings was presented for 3.33 sec, one above the other, and the spoken word for one of the drawings was played. Participants were required to respond by button press to indicate which drawing matched the spoken word. Each pair of drawings was presented twice, with a different drawing used as the target word, creating 108

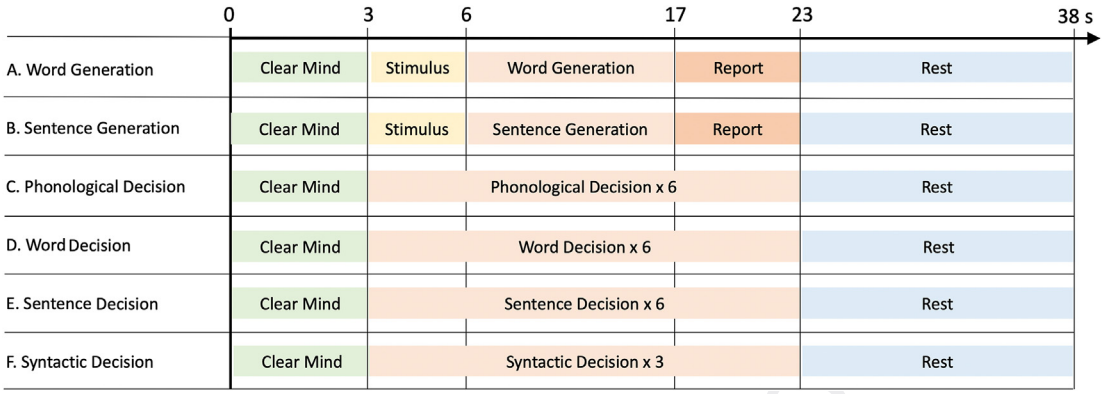

Fig. 2 – Timings within a single trial for the six tasks used with fTCD.

epochs in total. Epochs were presented in a pseudorandomised order, with no repetitions of a drawing pair in successive epochs. The same order was used for all participants. The target location was pseudorandomised so that the target was presented at the top / bottom of the screen (and therefore eliciting a left / right button press) in 50% of all epochs to avoid a response bias. In each fTCD trial, six epochs (drawing pairs) were presented, each lasting 3.33 sec. The pseudorandom order was designed so that within an fTCD trial, there were equal numbers of top or bottom targets. This ensured that odd-even split-half reliability data would not be affected by trial-to-trial variation in which hand is used. Participants were instructed to respond as quickly and accurately as possible.

2.5. *Computing laterality indices*

2.5.1. *Online battery tasks*

The Edinburgh Handedness Inventory was scored in the standard way, reflecting how often the left / right hands were used across all the items in the inventory. This score was converted into an index as described below. Indices greater than 0 were categorised as right-handed, and indices less than 0 as left-handed. The Porta Test classifies participants as being either left or right eye dominant.

For the Rhyme Decision and Word Comprehension tasks, each participant's RT for correct trials was used to calculate a laterality index that corresponds to a z-score, known as a LIz score. This can be readily derived from a t-test conducted on each participant's individual data, where accurate log response times (after outlier exclusion) are the dependent variable and visual field is the independent variable (see Parker et al., 2021). This estimates sensitivity to stimuli presented in either visual field and acts as a laterality index. The LIz score is very highly correlated with the more traditional Laterality Index, computed as (L-R)/(L+R), but it allows one to identify participants who show a statistically significant RT advantage for one side, i.e., where the LIz on a 2-sided test has $p < .05$.

For Dichotic Listening, a laterality index was calculated based on trials in which participants correctly identified one of the two consonant-vowel sounds. The count of correct responses that corresponded to each side was used to generate an accuracy laterality index. The index was calculated in the traditional fashion: 100 × (Left − Right) / (Left + Right). The laterality index allows us to relate our findings to prior research that uses this index. In addition, we computed a LIz-score for each participant, using the formula:

$$z = (pL-.5)/\mathrm{sqrt}[(pL.pR)/n]$$

where pR is the proportion of R responses, pL is the proportion of L responses, and n is the total L and R responses. As with the Rhyme Decision task, the z-score is highly correlated with the traditional laterality index, but has the advantage that it can be used to test whether an individual's lateral bias is unlikely to have arisen by chance.

We made two small modifications to the pre-registered plan; first we flipped the sign where necessary to ensure that left-hemisphere superiority was reflected in a positive score on all measures. This has no material effect on any computations, but gives better consistency with other research. Second, for the laterality z-score on Dichotic Listening, scores were censored at +/- 10, to avoid undue influence from a handful of extreme scores (participants who responded overwhelmingly to one ear).

2.5.2. *Functional transcranial Doppler ultrasound tasks*

Following Woodhead et al. (2019) we calculated a laterality index using a customised R script (R Core Team, 2016). The cerebral blood flow velocity (CBFV) data were first down-sampled from 100 to 25 Hz and then segmented into epochs. Spiking or dropout data-points were identified as being outside of the .0001–.9999 quantiles of the CBFV data. If only a single artefact data-point was identified within an epoch, it was replaced with the mean for that epoch. If more than one data-point was identified, the epoch was rejected. The CBFV was then normalised (by dividing by the mean and multiplying by 100) such that the values for CBFV become independent of the angle of insonation and the diameter of the middle cerebral artery. Heart cycle integration was then used to normalize the data relative to rhythmic modulations in CBFV. *The pre-registration document stated that "Epochs are baseline corrected using the interval from −10 to 0 sec pre-stimulus time (where the onset of the 'Clear' stimulus is used as the start of the trial, 0 sec pre-stimulus time)." Woodhead* et al. *used a shorter interval of -5 to 2 sec for baseline correction to avoid activity from the prior trial influencing the baseline. We had aimed to avoid that*

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

problem by having slightly longer epochs, but inspection of blood flow plots (see Fig. 6 below) showed this was not the case, and so we reverted to the baseline of -5 to 2 sec, consistent with the baseline period used by Woodhead et al. For completeness, all key analyses have been re-run using the original pre-registered baseline, to confirm this has minimal effect on outcomes (see Supplementary file 8 [Available at: https://osf.io/g9tqh/]).

Finally, artefacts were identified as values below 60% and above 140% of the mean normalised CBFV: any epochs containing such artefacts were rejected. The laterality index was then computed as the mean difference between blood flow velocity in left and right channels over a period of interest that is specified in advance for each task. For tasks without an overt speech 'report' stimulus (tasks C–F), the period of interest was from 6 to 23 sec peri-stimulus time to cover the whole period where the participant is performing the decision task; left-hemisphere blood flow typically increases as each item in the trial is responded to. For tasks with a 'report' stimulus (tasks A-B) the period of interest was from 6 to 17 sec to avoid capturing activity related to the overt speech production. The standard error of the laterality index for each individual was computed from the laterality index obtained across individual trials, and was used both to identify outliers (individuals with unreliable laterality indices) and to categorise individuals in terms of direction of laterality. When the 95% confidence interval of the laterality index spanned zero, laterality was coded as bilateral; otherwise, it was coded as left or right depending on the sign of the difference.

In the previous study by Woodhead et al. (2019) we excluded cases with fewer than 12 acceptable trials and then used the Hoaglin-Iglewicz (1987) criterion for outlier detection: this involves identifying cases that have values well outside the interquartile range for the group. This was used to identify individuals where the laterality index for a given task had an unusually large standard error, indicative of high trial-by-trial variation. However, we noted that some individuals had very low standard errors, despite having fewer than 12 useable trials, and so in the current study our preregistration specified a minimum number of 10 trials (out of 18 trials administered), while retaining the Hoaglin-Iglewicz method for removing data for a given subject and condition when the standard error of the laterality index was high.

2.6. *Sampling and Analysis plan*

The analysis starts with presentation of descriptive data, including distributions of scores by handedness, and split-half reliability of measures. The hypotheses are then tested, following the preregistration from Table 1 of our original Stage 1 report. The preregistered text is presented with each analysis. All analyses are conducted with alpha = .02 and power .9. A Rmarkdown script to run analyses on simulated data is available on Open Science Framework: https://osf.io/9dbrg/?view_only=357994fa8f6b49ee83964f5108d82ee2.

We report how we determined our sample size, all data exclusions, all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study.

3. Results

3.1. *Participants*

See Figure 3.

3.1.1. *Departures from pre-registration plan*

Our original flowchart had stated there were four online behavioural laterality tasks, but, as stated in the Stage 1 report, one of these had been dropped after pilot testing showed it had poor reliability, and only three tasks were therefore used. The original flowchart from the Stage 1 report was in error in mentioning 4 tasks, and this is corrected in Fig. 3.

Our plan had been to recruit 300 left-handers and 150 right-handers for the online behavioural battery, and from these to select 112 left-handers and 112 right-handers for in-person testing. Because of the Covid pandemic, the time periods when it was possible to test in person were greatly restricted, and testing had to be carried out under strict conditions to ensure safety (researchers wearing full personal protective equipment in a ventilated space, and cleaning of equipment between participants). Furthermore, researchers and participants could become unavailable for testing at short notice because of a positive Covid test or notification of a contact with an infected person, and some participants were understandably reluctant to attend in-person testing.

The disruption due to these factors meant we did not meet our target numbers for in-person testing, despite over-recruiting for online testing. We decided to relax the criteria for language competence of participants; rather than excluding people, it seemed preferable to include at Step 2 any participant who had completed Step 1 and was willing to come for testing, and then check retrospectively to see whether inclusion of these participants had any impact on the results. This gives a larger sample, which gives more confidence that null results are truly null.

To justify this approach, we considered whether either raw or absolute laterality indices were correlated with either of the language screening measures and showed they were not (See section 3.5 below and Supplementary file 6 (Available at: https://osf.io/g9tqh/)).

Fig. 3 presents the original preregistered recruitment flowchart modified to show actual numbers recruited. In total we tested 345 left-handers, 118 of whom were seen for fTCD assessment, and 276 right-handers, 112 of whom were seen for fTCD assessment.

Fig. 3 also shows the new rather than preregistered approach to the language screen. Group H (high proficiency) corresponds to cases who meet preregistered criteria: they are either native English speakers, or non-native speakers who passed the preregistered criterion for language competence (LexTALE of 80 or more, and no more than 3 errors on full Games with Words test). Group M (moderate proficiency) failed the language screen. In the original flowchart, the language screen was used to exclude those in group M prior to Step 2, whereas in the final study, we included these

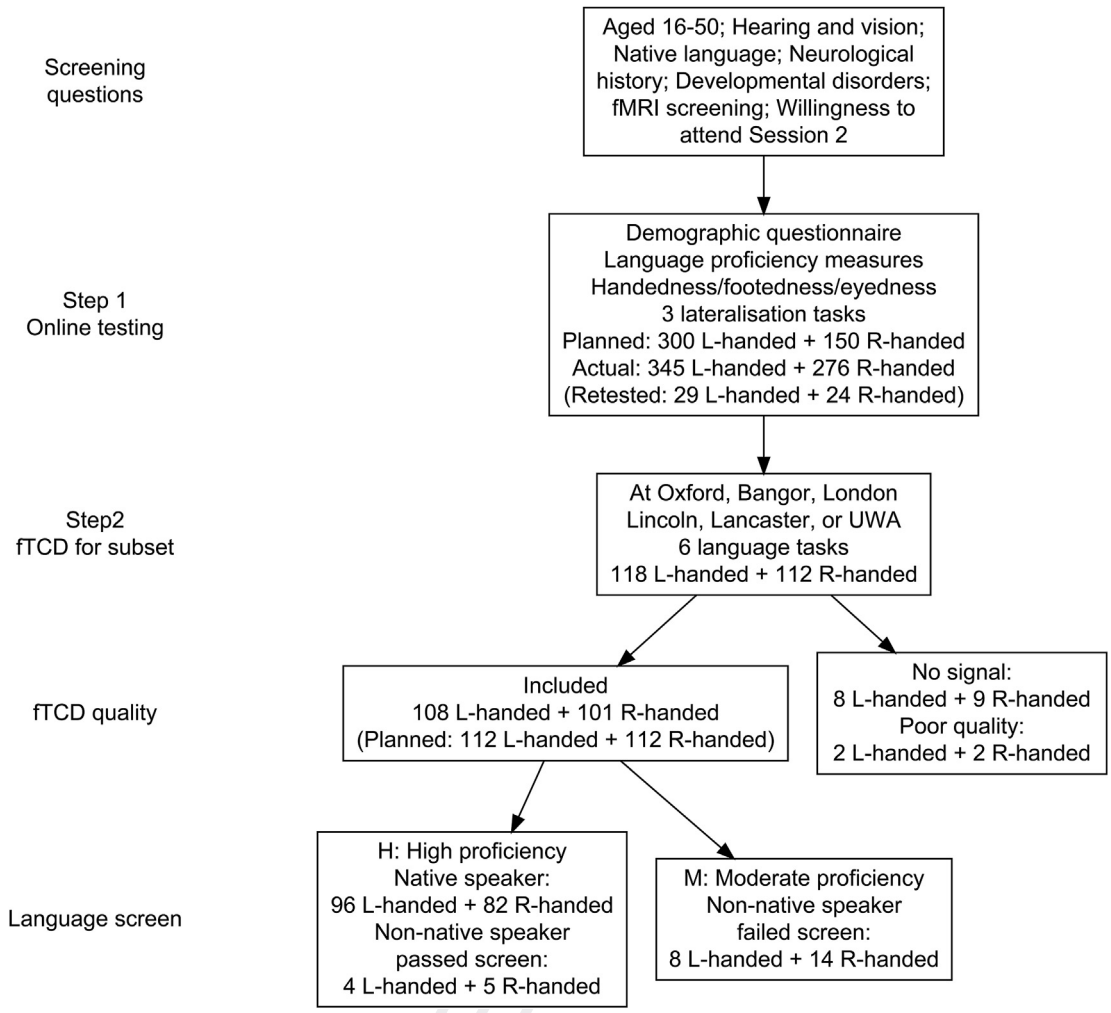

Fig. 3 – Participant recruitment flowchart.

individuals. In Supplementary file 3 (Available at: https://osf.io/g9tqh/) we compare the current results with those obtained using the original, stringent language cutoff (i.e., excluding Group M), and show that differences are minimal.

3.1.2. *Demographic data*

Table 1 shows demographic data for the subset of individuals tested on the online behavioural battery only, and the subset who also completed the session with fTCD. To obtain the total number completing the online session, the No FTCD data and With FTCD data columns should be summed. It is evident from inspection that differences between the two subgroups are not substantial. The median time difference between the online battery and the fTCD session was 61 days (range: 0–240 days).

3.1.3. *Subsample for test-retest study*

We had preregistered that we would do a test-retest reliability check on online behavioural tests for 50 participants. In practice, a subsample of 53 participants was retested on part of the online battery within 28 weeks of the first session (median = 73 days, range = 11–195 days). There were 15 left-handed females, 14 left-handed males, 13 right-handed females, and 11 right-handed males. The retest session included the Rhyme Decision and Word Comprehension tasks, which were new, but not Dichotic Listening, for which we had adequate evidence of test-retest reliability from the previous study by Parker et al. (2021).

3.2. *Step 1: online behavioural testing*

3.2.1. *Exclusions*

We adopted the same procedures as Parker et al. (2021) for excluding participants. On Dichotic Listening, 44 participants were excluded because their accuracy on 'same' trials (where both ears heard the same syllable) was less than 75%. On Rhyme Decision and Word Comprehension 61 and 20 participants respectively were excluded because overall accuracy was less than 75%. No exclusions were made on the basis of the size of ear difference, which was substantial in some cases.

3.2.2. *Derivation of laterality indices*

Visualisations of the raw data from which laterality indices were computed can be found in Supplementary file 7 (Available at: https://osf.io/g9tqh/). This Supplement also shows a

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

Table 1 – Demographic characteristics of sample.

	No FTCD data		With FTCD data	
	Left-Handed (N = 227)	Right-Handed (N = 164)	Left-Handed (N = 110)	Right-Handed (N = 103)
Gender				
Female	132 (58.1%)	86 (52.4%)	66 (60.0%)	60 (58.3%)
Male	94 (41.4%)	77 (47.0%)	41 (37.3%)	43 (41.7%)
Missing	1 (.4%)	1 (.6%)	3 (2.7%)	0 (0%)
Age (yr)				
Mean (SD)	28.7 (9.86)	25.9 (8.34)	28.0 (9.79)	25.8 (8.64)
Median [Min, Max]	25.0 [16.0, 50.0]	23.0 [16.0, 50.0]	24.0 [16.0, 50.0]	23.0 [17.0, 50.0]
Native English speaker				
No	27 (11.9%)	38 (23.2%)	14 (12.7%)	19 (18.4%)
Yes	200 (88.1%)	126 (76.8%)	96 (87.3%)	84 (81.6%)
Bilingual				
No	176 (77.5%)	112 (68.3%)	86 (78.2%)	77 (74.8%)
Yes	51 (22.5%)	52 (31.7%)	24 (21.8%)	26 (25.2%)
Edinburgh Handedness LI				
Mean (SD)	−77.5 (26.5)	82.9 (21.6)	−72.9 (26.1)	87.8 (18.4)
Median [Min, Max]	−86.7 [-100, 0]	89.5 [9.09, 100]	−80.0 [-100, 0]	100 [16.7, 100]

scatterplot showing how the LIz score relates to the conventional laterality index for Dichotic Listening. The principal difference is that the LIz follows the normal distribution, with a sigmoid shape at the extremes (truncated in the Figure S7.2 because of the censored scale (Available at: https://osf.io/g9tqh/)). For subsequent analyses, we use LIz, as this allows us to compare different tasks on a common scale.

3.2.3. *Distribution of laterality indices*

Before testing specific predictions about interrelationships between measures, we conducted preliminary analysis on LIz values for all three online tasks, to test for normality, to check for significant lateralisation in left- and right-handers, to compare laterality between handedness groups, and to compute split-half and test-retest reliability for laterality indices. For these, and subsequent analyses of handedness, a simple binary split into left- and right-handed was made according to whether the laterality index was above (R-handed) or below (L-handed) zero on the Edinburgh Handedness Inventory.

Fig. 4 shows distributions of scores for left- and right-handers as pirate plots, a form of beanplot with grey dots showing individual datapoints, the horizontal bar showing the mean, and the rectangle around the bar corresponding to the 95% Bayesian Highest Density Interval (Phillips, 2017). Table 2, distributions of LIz on the three tasks were non-normal, and the three tasks showed very different patterns of laterality. As expected from previous studies, on Dichotic Listening there was a clear right ear advantage in both left- and right-handers. Nevertheless, on the criterion of having an absolute LIz score of 1.96 or more, only 33.7% of participants were reliably left-lateralised, with 55.9% unlateralised, and 10.5% reliably right-lateralised. In addition, there was a small but statistically reliable difference between handedness groups, with stronger laterality in the right-handers. We did not assess test-retest reliability for this task, as we had done this in our previous study and found it to be high (Parker et al., 2021). Here we confirm excellent split-half reliability for this task.

The Rhyme Decision task was far less reliable, with split-half reliability [95% CI] of .41 [.33, .48] and test-retest reliability of .53 [.29, .72]. These figures indicate that laterality bias from zero on this test is well above chance, but there is a great deal of random variation. In addition, although the task showed statistically reliable laterality in both left- and right-handers in this large sample, the effect size was small, and

Table 2 – Descriptive statistics for three online laterality measures (LIz).

Statistic	Dichotic	Rhyme	Comprehension
N	324 LH + 253 RH	316 LH + 244 RH	334 LH + 267 RH
Mean (SD)	1.23 (3.15)	.32 (1.78)	−1.13 (2.38)
Skew	−.13 (p = .186)	−.46 (p < .001)	−.41 (p < .001)
Kurtosis	2.75 (p < .001)	1.50 (p < .001)	1.78 (p < .001)
Shapiro–Wilk normality	p < .001	p < .001	p < .001
Mean (SD): L-hander	.96 (3.26)	.28 (1.92)	−1.39 (2.41)
Mean (SD): R-hander	1.59 (2.96)	.38 (1.59)	−.81 (2.31)
one-group t: L-hander	t = 5.3; p < .001	t = 2.6; p = .011	t = −10.5; p < .001
one-group t: R-hander	t = 8.5; p < .001	t = 3.7; p < .001	t = −5.7; p < .001
R-hander *vs* L-hander t	t = 2.4; p = .008	t = .7; p = .247	t = 3.0; p = .001
Split-half r_s [95% CI]	r_s = .66 [.60, .71]	r_s = .41 [.33, .48]	r_s = .62 [.56, .67]
Test-retest r_s [95% CI] (N = 53)		r_s = .53 [.29, .72]	r_s = .55 [.31, .73]

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

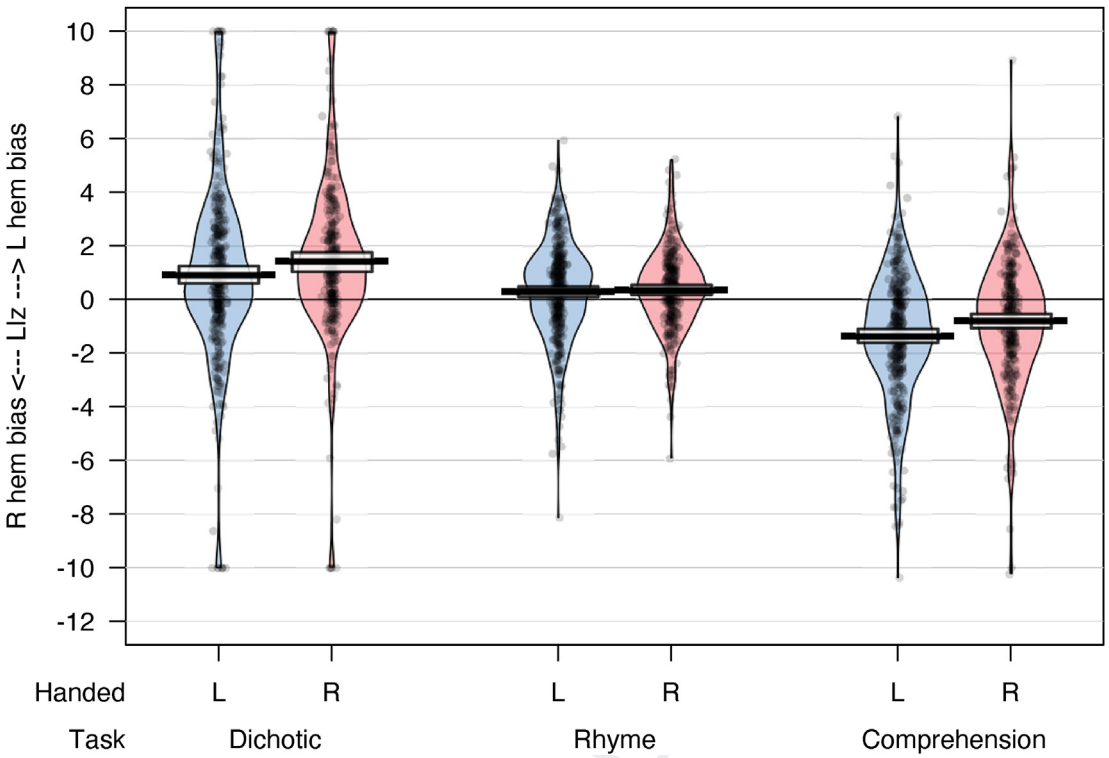

Fig. 4 – Pirate plot distributions of LIz scores on online tasks.

most individuals were not significantly lateralised on the criterion of having an absolute LIz score of 1.96 or more: 13.8% of participants were reliably left-lateralised, with 77.3% unlateralised, and 8.9% reliably right-lateralised. Furthermore, there was no effect of handedness on laterality on this task.

The Word Comprehension task did rather better than the other tasks in terms of reliability, with split-half reliability of .62 [.56, .67] though test-retest reliability was lower at .55 [.31, .73]. The striking observation about this task was that it showed a laterality bias in the opposite direction to what is usually seen in language tasks. Responses were faster when the target picture that matched the auditorily presented word occurred in the left visual half-field, which projects directly to the right hemisphere. On the criterion of having an absolute LIz score of 1.96 or more: 6.4% of participants were reliably left-lateralised, with 62.3% unlateralised, and 31.2% reliably right-lateralised. Furthermore, there was a significant effect of handedness, with the laterality index being more negative in left-handers than in right-handers. We consider the implications of these findings in the Discussion below.

3.2.4. *Testing prediction 1: Fit of two-factor model to behavioural data*

Prediction 1 stated: ***The pattern of correlation between laterality indices from online measures will reflect the extent to which they involve implicit speech production, rather than whether they involve spoken or written language. Thus we anticipate dissociation between the rhyme judgement task and the other two measures (Dichotic Listening and Word Comprehension task), which is not accountable for in terms of low reliability of measures.***

Departure from pregistration: We had planned to do a formal comparison of model fit using AIC weights, but we realised our data were inadequate for this because our Model A was, in formal terms, just-identified: it simply estimated three pairwise correlations from the data, and always gave perfect fit, regardless of the size or direction of correlations. We considered alternative approaches to the analysis, but decided to just report the correlations at this stage, as the pattern of results was distinctive, and we had already planned to incorporate the online behavioural measures into the SEM analysis that includes the fTCD measures (see section 3.4 below).

Fig. 5 shows scatterplots of the bivariate relationships between the three variables. Spearman correlations are shown with 95% confidence intervals estimated using the *spearman.ci()* function from the RVAideMemoire package (version .9–79; Hervé, 2021) with 1,000 iterations. It is evident from inspection that we can reject model C, in which all three LIs are independent, and model B1, where only Dichotic Listening and Rhyme Decision are correlated. The strongest correlation is between the two visual tasks, Rhyme Decision and Word Comprehension, as predicted by model B2.

Note that correlations will be influenced by test reliability. Indeed, the correlation between Rhyme Decision and Word Comprehension is close in magnitude to the split-half reliability of the two measures. An estimate of the association between these measures after adjusting for the split-half reliabilities can be obtained using the Spearman-Brown correction for attenuation, r.xy(corrected) = r.xy(observed)/sqrt(r.xx * r.yy), which gives a value of .818.

Evaluation of Prediction 1: Separable dimensions for receptive and production online laterality tasks. This prediction was not confirmed. The baseline hypothesis of a single laterality factor was also rejected.

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

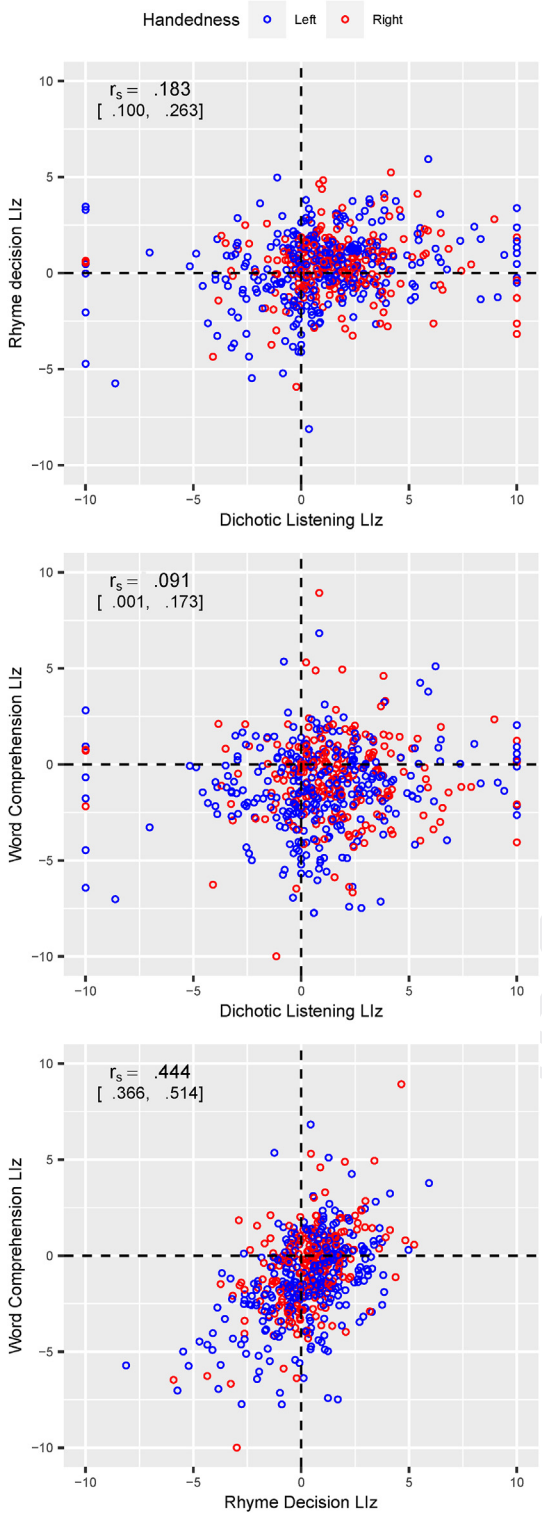

Fig. 5 – Bivariate distributions of LIs on behavioural tasks.

3.3. *Step 2: Functional transcranial Doppler ultrasound (fTCD)*

We excluded 4 participants who met our criteria for outliers on two or more fTCD tasks. For the remaining 209 participants, the numbers with fewer than 10 useable trials on the six tasks (A = Word Generation, B = Sentence Generation, C = Phonological Decision, D = Word Decision, E = Sentence Decision and F = Syntactic Decision) were 1, 3, 2, 6, 0 and 3 respectively.

3.3.1. *Cerebral blood flow velocity*

Fig. 6 shows the mean time course of blood flow velocity on left and right channels for left- and right-handers. In addition, the difference between left and right channels is plotted in black, after adding 100 to the values so that they can be shown on the same plot (with scale on right side axis). The laterality index is computed as the mean difference score (shown in black) over the period of interest. For Word Generation and Sentence Generation tasks, the LI is computed during a period corresponding to silent generation; the waveform peaks again after this period, corresponding to the activity from the subsequent spoken response. For the other tasks, a series of items is presented in each trial and no spoken response is required. The periodic fluctuations in the response correspond to the individual items that are responded to.

Inspection of this figure indicates that we see a strong left hemisphere bias in both handedness groups for Word Generation, Sentence Generation and Phonological Decision, whereas the other tasks do not show this pattern.

3.3.2. *Behavioural performance on FTCD tasks*

The mean number of words produced per trial was 4.45, (SD = .82) for Word Generation and 9.47, (SD = 1.23) for Sentence Generation. For the four decision tasks, accuracy was recorded as mean percentage correct: 88.16 (SD = 7.29) for Phonological Decision, 98.96 (SD = 1.43) for Word Decision, 87.36 (SD = 6.34) for Sentence Decision, and 81.43 (SD = 9.88) for Syntactic Decision.

3.3.3. *Laterality indices*

The pirate plot in Fig. 7 shows LI values for the six tasks (A = Word Generation, B = Sentence Generation, C = Phonological Decision, D = Word Decision, E = Sentence Decision and F = Syntactic Decision) for left- and right-handed participants.

Table 3 shows basic statistics for the fTCD laterality indices, in the same format as for the online tasks. Right-handers showed significant left-lateralisation on Word Generation, Sentence Generation, Phonological Decision, and Sentence Decision, but were not lateralised for Word Decision or Syntactic Decision. Left-handers were significantly left-lateralised for Word Generation, Sentence Generation and Phonological Decision, were not lateralised for Sentence Decision or Syntactic Decision, and were significantly right-lateralised for Word Decision. The direct comparison between left- and right-handers showed significantly greater left-lateralisation in right-handers on all tasks except Word Decision and Syntactic Decision.

All tasks had split-half reliability coefficients of .72 or above, except for Word Decision, where the coefficient was only .52.

Shapiro Wilk tests revealed significant non-normality for Sentence Generation, Phonological Decision, Word Decision and Sentence Decision, though values of skewness and kurtosis were generally not extreme.

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

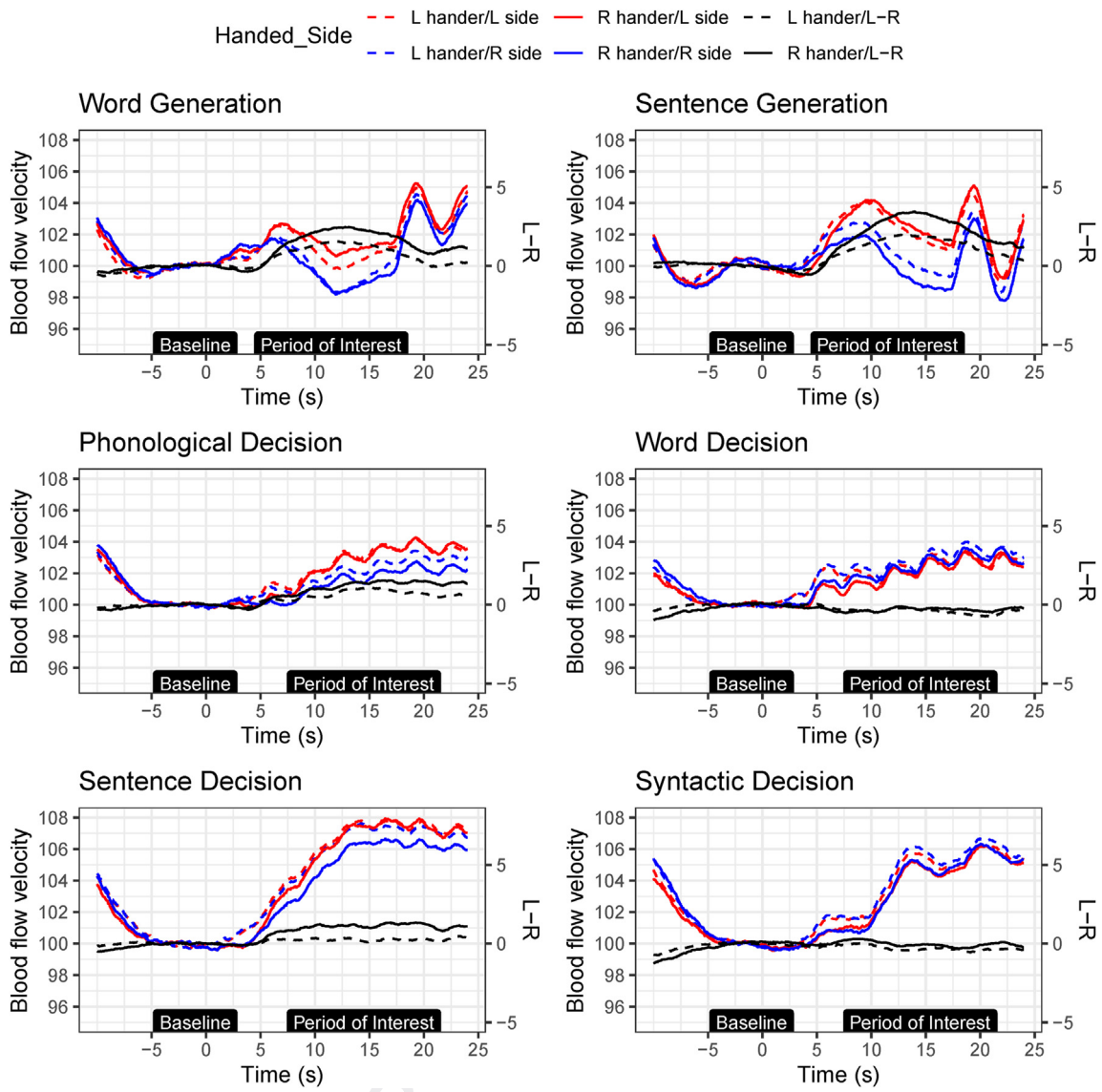

Fig. 6 – Timecourse of left and right hemisphere blood flow velocity (left axis) and L-R difference (right axis) for six tasks in left- and right-handers).

As noted above, there were a few participants with missing data on a single measure. Before running the SEM analysis, the *mice* package (Van Buuren & Groothuis-Oudshoorn, 2011) was run with default settings in R to impute these missing values.

3.3.4. *Testing prediction 2: Fit of two-factor model to FTCD data*

Our second preregistered prediction was: ***The data will fit a model where "language generation" tasks cluster together on one factor, and "receptive language" tasks on a second factor.*** It was further predicted that factors will be correlated, but the fit of a 2-factor model will be superior to a single-factor model where all LIs load on a common factor.

The analysis conducted by Woodhead et al. (2019, 2020) used an exploratory bifactor model in which each task could load on each of two factors. Because there were two measures for each task (from test and retest sessions), this exploratory approach was adequately powered. In the current study, two tasks differed to Woodhead et al.'s studies (List Generation and Semantic Decision were removed, and Word Generation and Word Decision were added), and we only had one measurement occasion for each of the six measures. Accordingly, we used confirmatory factor analysis, using a prespecified two-factor model that constrains which indicators can load on two factors. This was compared to a unitary model, in which all tasks load on a single factor.

Fig. 8 shows the pattern of correlations between LIs for the different tasks as a heatmap, with values for left-handers above the diagonal and those for right-handers below. Pearson correlations are shown here, as these relate more directly to the analysis of covariances that is the basis of Structural Equation Modeling. The two-factor model predicts that correlations will form two clusters, with positive correlations within the first three tests, and within the last three tests, but weaker or absent correlations across these two clusters of measures. In both handedness groups, the heatmap shows

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

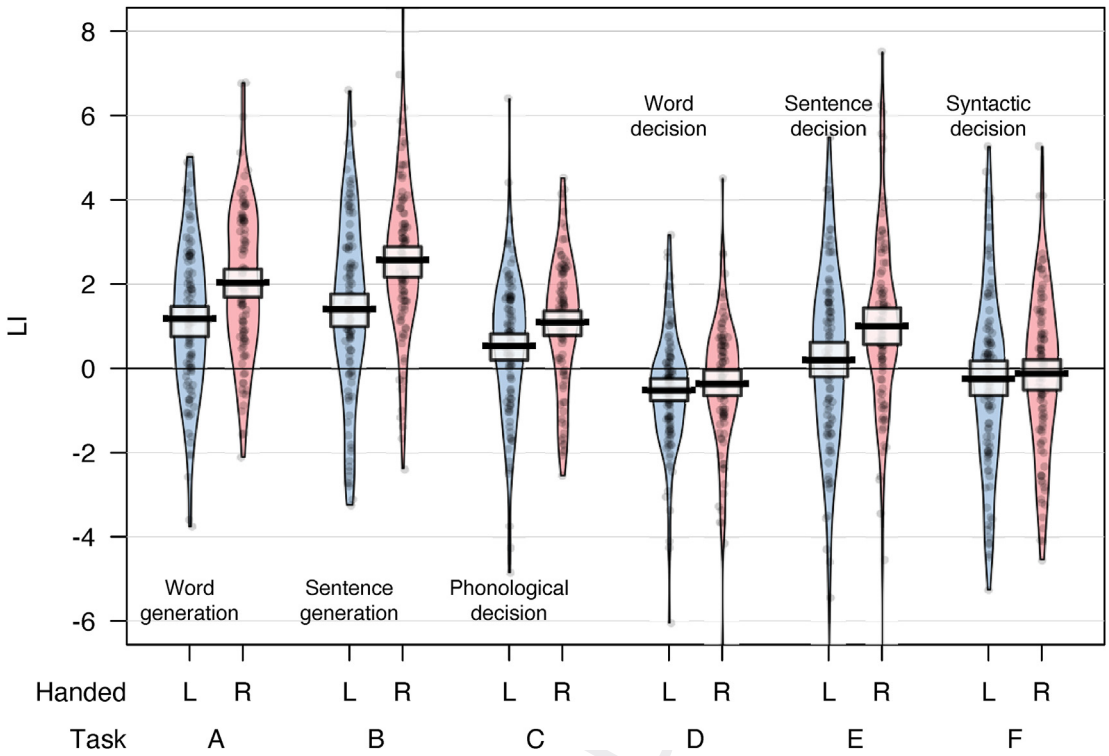

Fig. 7 – Distributions of fTCD LIs on six tasks for 104 left-handers and 91 right-handers.

Table 3 – Descriptive statistics for six fTCD laterality indices.

Statistic	A	B	C	D	E	F
N	108 LH + 99 RH	107 LH + 99 RH	105 LH + 100 RH	104 LH + 96 RH	105 LH + 99 RH	105 LH + 99 RH
Mean (SD)	1.59 (1.85)	1.95 (2.05)	.82 (1.65)	−.40 (1.42)	.59 (2.29)	−.20 (2.04)
Skew	−.03 (p = .860)	−.19 (p = .274)	−.31 (p = .066)	−.34 (p = .047)	−.43 (p = .013)	.04 (p = .819)
Kurtosis	.04 (p = .914)	.35 (p = .309)	.89 (p = .009)	2.24 (p < .001)	2.13 (p < .001)	.01 (p = .983)
Shapiro–Wilk normality	p = .890	p = .081	p = .014	p < .001	p < .001	p = .290
Mean (SD) L-hander	1.18 (1.82)	1.36 (2.09)	.56 (1.70)	−.43 (1.32)	.20 (2.27)	−.29 (2.17)
Mean (SD) R-hander	2.04 (1.78)	2.58 (1.80)	1.10 (1.55)	−.36 (1.54)	1.01 (2.25)	−.10 (1.90)
one-group t L-hander	t = 6.8; p < .001	t = 6.7; p < .001	t = 3.3; p = .001	t = −3.4; p = .001	t = .9; p = .358	t = −1.4; p = .178
one-group t R-hander	t = 11.4; p < .001	t = 14.2; p < .001	t = 7.1; p < .001	t = −2.3; p = .023	t = 4.5; p < .001	t = −.5; p = .600
R-hander *vs* L-hander t	t = 3.4; p < .001	t = 4.5; p < .001	t = 2.4; p = .009	t = .3; p = .366	t = 2.5; p = .006	t = .7; p = .257
Split-half r_s [95% CI]	r_s = .76 [.69, .81]	r_s = .80 [.73, .85]	r_s = .72 [.63, .79]	r_s = .52 [.39, .63]	r_s = .83 [.76, .87]	r_s = .79 [.71, .85]

moderate correlations within both clusters of measures, and generally lower correlations across clusters, but there are some exceptions. Notably, there is a moderate correlation between Phonological Decision and Sentence Decision, which was not predicted by the two-factor model.

3.3.4.1. Structural equation modeling. Because Structural equation modeling (SEM) is not widely used in laterality research, we provide here a brief explanation, to aid interpretation of the subsequent analysis.

Structural equation modeling (Kline, 2011) is a method that allows a formal test of adequacy of competing models for explaining patterns of association between variables. The underlying assumption of this approach is that observed variables can be treated as indicators of underlying, unobserved latent variables.

Associations between latent factors and observed variables are shown in a path diagram, with latent factors in circles, and observed variables in boxes. Single-headed arrows indicate causal paths, and double-headed arrows indicate variances. Although means can be incorporated in SEM (and we shall be doing this in our analysis), the main use of SEM is to analyse patterns of covariances. Path diagrams have a precise mathematical interpretation, and can be converted into linear equations that specify the covariances between observed variables. Thus it is possible to obtain a measure of goodness of fit for observed data in relation to a model by comparing how far the observed covariances agree with those predicted by the model. We can already see by inspecting the heatmap of Fig. 8 that a single-factor model is unlikely to provide a good fit to the observed data, because it would not predict the clustering of correlations that is evident.

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

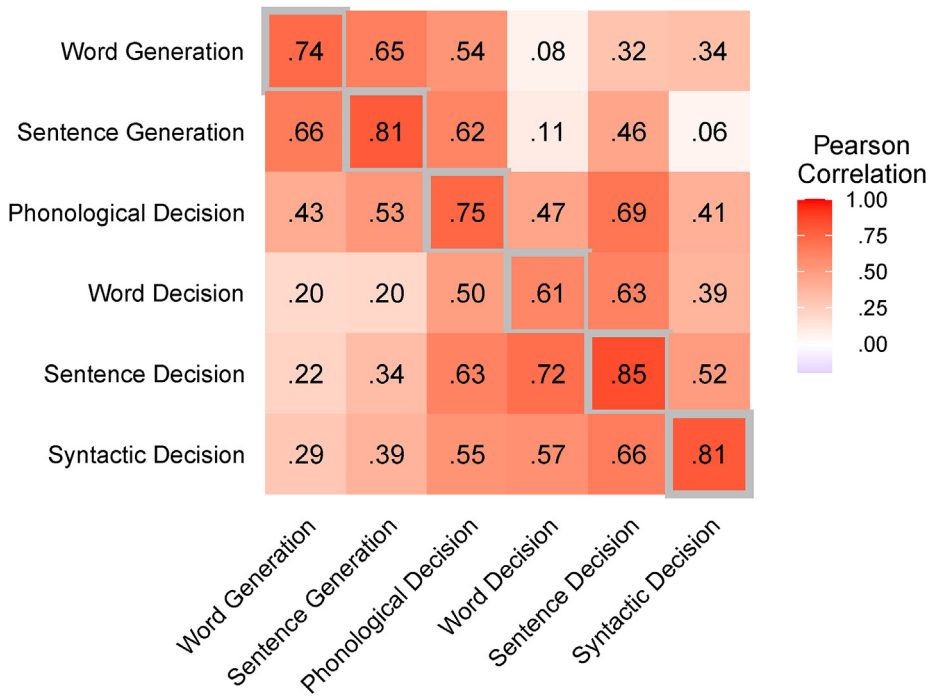

Fig. 8 – Heatmap showing correlations between laterality indices from six fTCD tasks, with values for L-handers above the diagonal, R-handers below the diagonal, and split-half reliability on the diagonal (cells with grey frame).

SEM does not arrive at a single algebraic estimation of model fit, but rather uses a maximum likelihood approach, whereby values for the paths from the factors to the observed variables are first assigned starting values, and the expected covariances between variables are computed with these values, and then compared to observed covariances. This process is iterated many times with different path estimates, with an algorithm adjusting paths on each run to reduce the mismatch between observed and expected (i.e., model-implied) covariances.

Path diagrams are shown below in Fig. 9 (single factor model), and Fig. 10 (two-factor model). **The two-factor model is equivalent to Fig. 4 from the preregistered document**. These include one path to each factor shown as a dotted line. This is a fixed parameter, set to 1, which is necessary to scale the estimates. Computationally, it makes no difference to the solution which path is fixed: for each factor this can be either one of the paths from an observed variable, or the variance of the factor. All the other paths are free to vary, and the estimation process will consider different values, to converge on a solution that gives the best fit. Some paths may have little impact on the solution, and may be dropped without any deterioration of fit.

There is no single method for evaluating the fit of a model to observed data (Schermelleh-Engel, Moosbrugger, & Müller, 2003). A χ^2 test can give an estimate of the extent of departure of observed values from expectation: a good model is one where χ^2 is small and has a high associated p-value, indicating that any difference between expectation and observation is likely to just reflect sampling error. It is usually possible to improve the fit by including additional paths or factors in a model until good fit is achieved, but this does not mean that the model is better: the goal is rather to obtain a parsimonious and theoretically meaningful model that does not include arbitrary parameters that are specified solely to fit the data. Note, however, that values of χ^2 are dependent on sample size, and with small samples, a small χ^2 value may suggest good fit, when differences between observed and model-implied covariances are large; in effect small samples may lack power to detect departures from model predictions, whereas large samples risk finding significant departures from perfect fit on the basis of trivial mismatch.

For this reason, a range of different measures of model fit has been devised. The first is the Root Mean Square Error of Approximation (RMSEA), a measure of approximate fit in the population that is largely independent of sample size. RMSEA is a measure of "badness of fit", where a value of zero indicates good fit. The Standardized Root Mean Square Residual (SRMR) considers the average size of fitted residuals after the model is fitted, and also ranges from 0 to 1. For both RMSEA and SRMR, values below .05 are generally regarded as indicating good model fit (Schermelleh-Engel et al., 2003).

Other indices have been developed that penalise models with a large number of parameters. The Comparative Fit Index (CFI) measures relative improvement of fit of a model relative to a model that assumes independence of all variables. CFI values of .95 or more are conventionally regarded as indicating acceptable fit. The Tucker–Lewis Index (TLI) similarly compares chi square of the observed model with an independence model, with values of .97 or more indicating good model fit (Schermelleh-Engel et al., 2003).

Models are 'nested' when a simple model can be derived from a more complex model by fixing at least one free parameter in the complex model. In Figs. 9 and 10, the single-factor model is equivalent to the two-factor model if the covariance between F1 and F2 is fixed to one. In such cases,

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

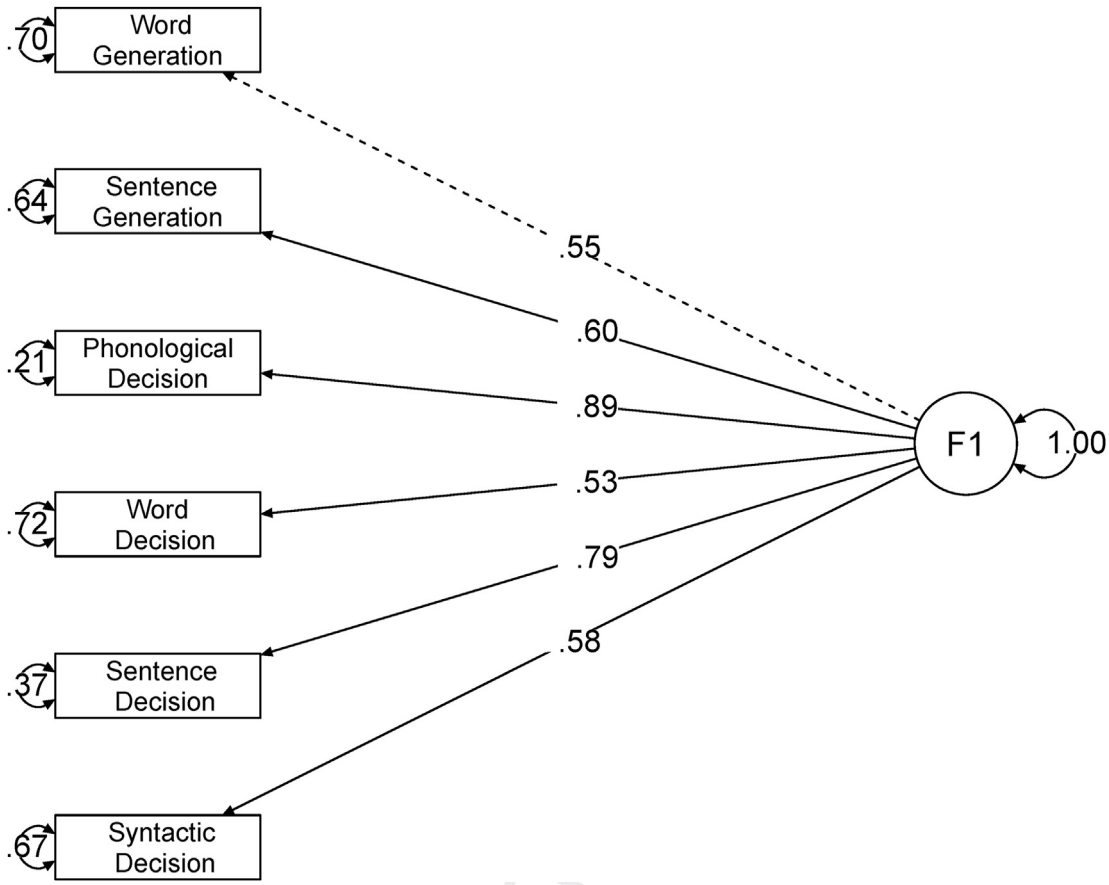

Fig. 9 – Single-factor model, showing standardized path coefficients obtained in the current analysis.

model fit can be compared by subtraction of the χ^2 and degrees of freedom for the two models; the difference in χ^2 is then evaluated; if it is nonsignificant, this indicates that the simpler model gives as good a fit as a complex model with more parameters, in which case the simpler model is preferred.

Following recommendations by Schermelleh-Engel et al. (2003) we report here values for χ^2, CFI, TLI, RMSEA and SRMR. Detailed outputs for all SEM analyses are available in Supplementary document 8 (Available at: https://osf.io/g9tqh/).

We used the lavaan() package (Rosseel, 2012) to perform the preregistered model comparison. To take into account non-normality of some variables, the WLSMV estimator was specified; this uses weighted least squares with robust standard errors and a mean- and variance adjusted test statistic, and makes no distributional assumptions about the observed variables. When this estimator is used, robust χ^2 values should be used to evaluate model fit, though the unadjusted χ^2 values are used for model comparison. Table 4 summarises the main output of the model-fitting. The fit of both the single-factor and the two-factor model is poor.

Therefore, as planned, we divided the sample into two random subsamples, 1 and 2. The first subsample was used in an exploratory analysis, based on that used by Woodhead et al. (2021), and the second for cross-validation. Fuller results from the analysis are given in Supplementary material (Available at: https://osf.io/g9tqh/). Because we had data from a single session, for both exploratory and cross-validation analyses, we used the LIs from the odd and even trials to give two indicators per task. We started with a model with two factors, where all 12 measures (2 measures from 6 tasks) were allowed to load on both factors, except for Sentence Generation. This was an indicator variable with a loading of 1 on Factor 1, and no loading on Factor 2. To ensure model identification, the variance of Factor 1 was free to vary, and variance of Factor 2 was set to 1. Although this model converged with good fit, there were warnings indicating problems with unfeasibly small eigenvalues. However, when non-significant paths were dropped from the model (from Word Decision and Syntactic Decision to Factor 1, and from Word Generation to Factor 2), there was good model convergence with plausible parameters and excellent fit (CFI = 1 and RMSEA = 0). This same model was then evaluated with the hold-out sample, and again the fit was good. We therefore took this model forward to the next stage of analysis, first checking the fit with the original full sample and with the original LIs based on all trials. The path diagram is shown in Fig. 11 and summary output is shown in Table 4; the fit was a significant improvement on the fit of the original 2-factor model.

Evaluation of Prediction 2: Separable factors for receptive and production fTCD laterality measures

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

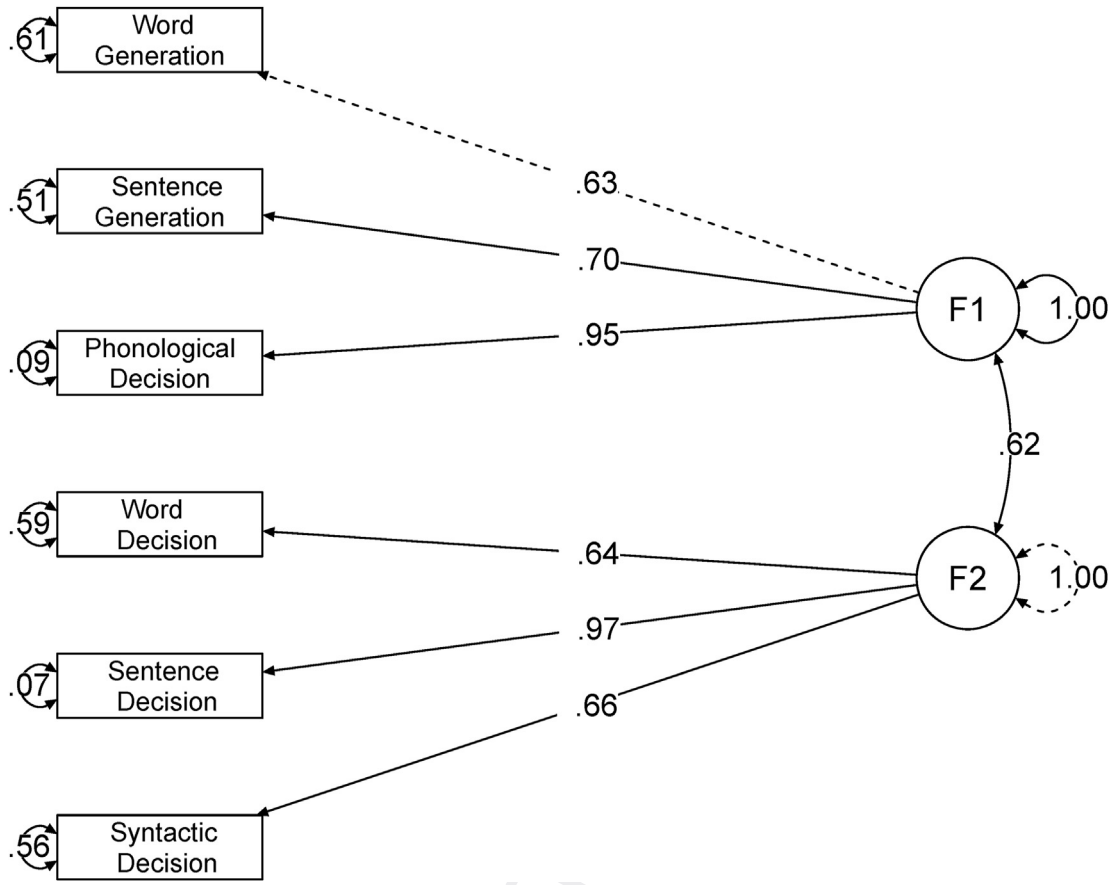

Fig. 10 – Two-factor model, showing standardized path coefficients obtained in the current analysis.

Table 4 – Fit statistics for 1-factor (Model.1F), 2-factor (Model.2F) and modified 2-factor (Model.2Fn) models (N = 209).

Estimate	Model.1F	Model.2F	Model.2Fn
CFI	.92	.97	1.00
TLI	.87	.94	1.01
SRMR	.12	.08	.04
RMSEA [95% CI]	.13 [.09, .17]	.08 [.04, .13]	.00 [.00, .08]
χ^2	39.99, $p < .001$	19.58, $p = .010$	4.92, $p = .550$
robustχ^2	79.00, $p < .001$	43.65, $p < .001$	14.09, $p = .030$
DF	9	8	6

This prediction was not confirmed. The baseline hypothesis of a single laterality factor was also rejected. A modification of the two-factor model that allowed additional paths from Phonological Decision to Factor 2, and from Sentence Decision to Factor 1 gave a good fit.

3.3.5. *Testing prediction 3: Model equivalence for left- and right-handers*

The third prediction was: better model fit will be obtained when different parameters are estimated for left- versus right-handers, compared with when all parameters are equated for the two handedness groups.

The approach we adopted is a standard one used when structural equation modeling (SEM) is applied to evaluation of measurement models in other domains, where it is described as a test of measurement invariance. Essentially, the data from left- and right-handers are analysed together in a series of nested models; these pose increasingly stringent constraints on which parameters of the model are allowed to vary for the two handedness groups. Detailed outputs of the nested models are given in Supplementary document 8 (Available at: https://osf.io/g9tqh/).

Initially, a model is fit in which all the paths, covariances, and intercepts are free to differ between left- and right-handers. This model is tested against a model of 'metric invariance', which sets the loadings from each observed variable to the factors to be the same for the two groups. If the fit of the model does not worsen, we can assume the basic model structure is equivalent for the two groups. This test of equivalence was passed (see Table 5).

At the next step, (scalar invariance), the item intercepts are set to be the same across groups. Once again, the model fit did not worsen (see Table 5).

Previously, Woodhead et al. (2021) had found weaker covariance between factors in left-handers than in right-handers. To test whether this was the case for the current dataset, we added a further constraint, which was that the covariance between factors should be the same for the two groups. The covariance between factors was numerically

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

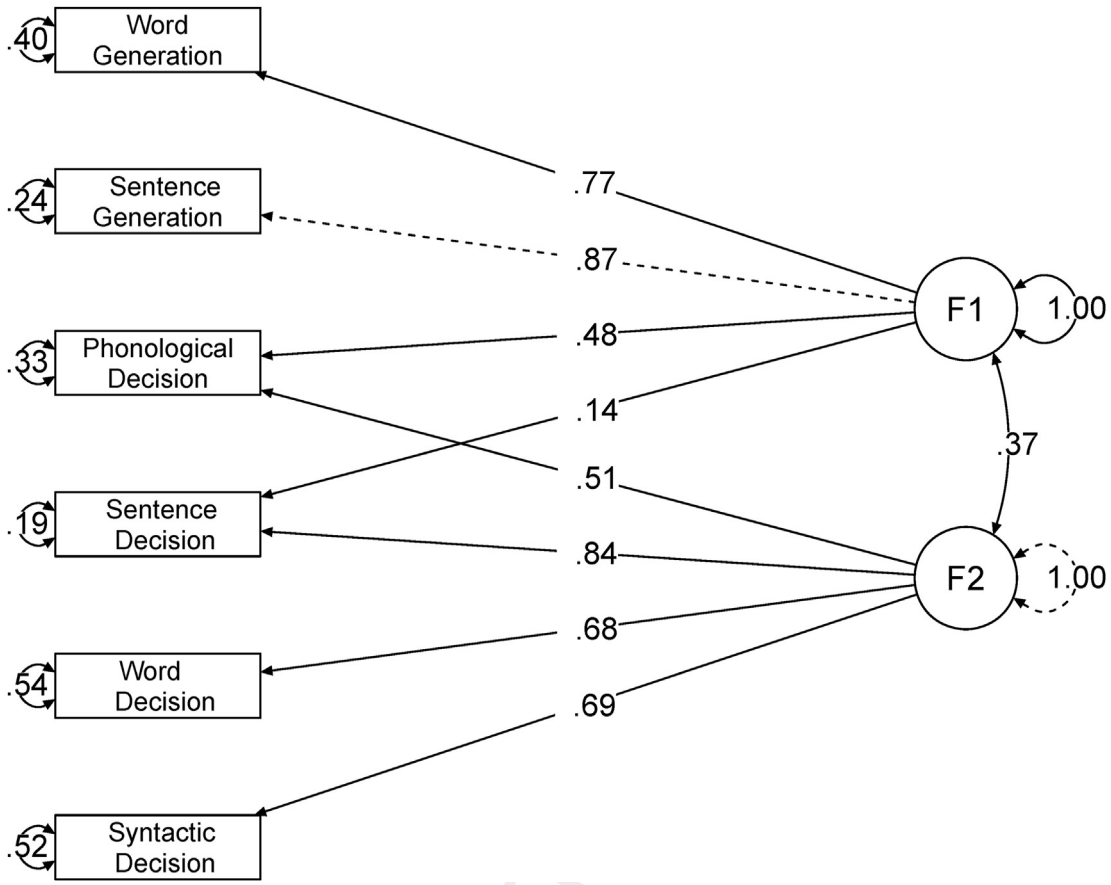

Fig. 11 – Revised two-factor model, showing standardized path coefficients obtained in the final analysis.

Table 5 – Nested tests of model equivalence for left- and right-handers.

Model	Group constraints	Df	χ^2	χ^2 diff	Df diff	p
1	None	12	8.23			
2	Equal loadings	19	12.87	5.80	7	.564
3	2 + Equal intercepts	23	14.00	2.99	4	.559
4	3 + Equal factor covariance	24	14.24	.06	1	.806
5	4 + Equal factor means	26	55.44	17.00	2	.000

lower in left-handers (.55, SE = .25) than right-handers (.64, SE = .23), but the difference was not large, and model fit was not impaired when the covariance was forced to be identical in the two handedness groups.

In a final step (strict invariance), we constrain item residual variances as well as factor loadings and intercepts to be equal across groups. Here we obtained a substantial worsening of model fit, indicating that the mean difference between groups on the latent factors is not the same.

In sum, results from the measurement invariance test showed that, contrary to our prediction, the same underlying structural model can be assumed to apply for both left- and right-handers, with the differences between handedness groups being explained solely in terms of differences in factor means, rather than in the pattern of covariances between the six LIs.

In a further exploratory analysis, we compared left- and right-handers on mean factor scores. Values for Factor 1 were: Left-handers mean = −.46, SD = 1.71; Right-handers, mean = .49, SD = 1.50. For Factor 1, the effect size for handedness (Cohen's d) was .59, and a t-test gave t (206.1) = −4.27, $p < .001$. For Factor 2, Left-handers mean = −.11, SD = .90; Right-handers, mean = .12, SD = .95. The effect size for handedness (Cohen's d) was .25, and a t-test gave t (204.2) = −1.78, $p = .077$.

Evaluation of Prediction 3: Different model parameters for left- and right-handers Our interpretation of this analysis was rather complex, given the range of possible results. Specifically, in the preregistration we stated: "***The simplest result would be to confirm structural invariance, i.e., no difference in model parameters for left- versus right-handers. This is unlikely, given our prior results, but it would indicate that handedness is unrelated to language laterality profile. On the basis of prior results we anticipate the best-fitting model will require different factor means for left- and right-handers. If so, we will ask whether specifying different means is sufficient to explain group differences***

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

– *this would indicate we can conceptualise the effect of handedness in terms of a population mean shift. Our simulated data suggests we may also need to specify different residuals for the two groups, reflecting greater variance in left-handers; if confirmed, this would indicate that a mean shift is insufficient to explain handedness effects, and suggest the underlying laterality distribution may contain a mixture of left- and right-biased individuals*. The obtained result clearly indicated that there were different factor means for left- and right-handers, and this was sufficient to account for group differences. The handedness difference was striking for Factor 1, but more marginal for Factor 2. Thus the most parsimonious account of the data was that handedness differences could be explained solely in terms of a shift away from left-hemisphere bias in left-handers for a laterality factor that had loadings from language generation tasks.

3.3.6. *Testing prediction 4: Categorical analysis of laterality indices*

Prediction 4 was: On categorical analysis, individuals who depart from left-brained laterality on one or more tasks will be more likely to be left-handed than those who are consistently left-lateralised.

The analysis so far has treated laterality as a continuum, but this continuum does have a zero-point, and negative scores indicate right-lateralisation and positive scores left-lateralisation. There are theoretical reasons to suppose that brain function might be influenced more by consistency in direction of lateralisation, than by degree. Thus, regardless of how strong or weak a laterality index is, brain functioning might be more efficient if all language functions are predominantly mediated by the same hemisphere.

As stated in our preregistration: **we first adopt the simple approach of dichotomising laterality at a cutoff of zero for each task, and then perform a** χ^2 analysis to test for association with handedness. For 6 measures, we adopt a Bonferroni-corrected alpha level of .02/6 = .003.

Results are shown in Table 6. The trend is similar for all six tasks, with the proportion who are left-lateralised averaging at 11% lower in left-handers than in right-handers, regardless of the mean LI for the task. The difference ranged from 1% for Syntactic Decision to 18% for Sentence Decision, but did not meet our prespecified significance criterion for any measure.

Our preregistered analysis plan stated: *If we find no significant difference between handedness groups, then we can conclude that any true difference in the percentage of left-lateralised individuals is 17% or less, i.e., much lower than the estimates of left-sided language lateralisation for left- and right-handers from lesion or Wada studies (67% and 95% respectively, a difference of 28%) We anticipate that all measures will differentiate left- from right-handers with our sample size of 224; any measure that does not do so would be regarded as an exception to the general rule that handedness is weakly predictive of language lateralisation*. Clearly, this prediction is not supported by the observed data (albeit with a smaller sample size).

We had preregistered a subsidiary analysis as follows: *After testing associations for individual measures, we will categorise individuals as either consistently left-lateralised on all tests, or right-lateralised on one or more tests, and conduct a* χ^2 test contrasting the proportion of left- and right-handers on this composite measure.

The proportions of left- and right-handers who are left-lateralised on between 0 and 6 tasks, using the same cutoff of zero, is shown in Table 7.

It is evident from Table 7 that a minority of individuals is consistently left-lateralised on all six tasks, regardless of handedness. The trend is for more right-handers to show this pattern than left-handers, but this difference is not significant on χ^2 test, $\chi^2 = 2.17$, $p = .141$. However, two of the tasks included in this analysis, Word Decision and Syntactic Decision, were not left-lateralised at the population level, and it could be argued they would just add noise to the analysis, which was intended to identify those who departed from the typical pattern of left-lateralisation. We therefore added an exploratory analysis, in which we excluded these two tasks. When only Word Generation, Sentence Generation, Phonological Decision and Syntactic Comprehension were considered, 43.7% of left-handers and 59.2% of right-handers were consistently left-lateralised.

It is noteworthy that this more categorical analysis finds rates of "atypical", i.e., non-left, lateralisation on language tasks that are task-dependent, and are lower than typically observed when methods such as Wada test or fMRI are used. This is the case even for the most lateralised task, Sentence Generation, where 91% of right-handers versus 76% of left-handers were left-lateralised.

This raises the question of how reliable the categorisation of laterality is with fTCD. In response to a reviewer suggestion, we recategorised participants using just the odd or even trials on these tasks, making it possible to identify cases where lateralisation as left or right was inconsistent. The percentages with inconsistent laterality, when a binary divide was

Table 6 – Proportions showing percentages of left lateralisation on fTCD tasks, with χ^2 test for L versus R hander difference.

Task	L hander %	R hander %	R-L hander %	χ^2	p
Word Generation	75.0	86.1	11.1	4.07	.044
Sentence Generation	75.9	91.1	15.2	7.55	.006
Phonological Decision	68.5	76.2	7.7	.78	.377
Word Decision	30.6	40.6	10.0	2.06	.151
Sentence Decision	55.6	73.3	17.7	6.25	.012
Syntactic Decision	48.1	49.5	1.4	.00	.950

Table 7 – Proportions of left- and right-handers with between 0 and 6 tasks left-lateralised on fTCD.

N tasks L lateralised	L-handers	R-handers
0	.059	.000
1	.059	.062
2	.196	.104
3	.137	.156
4	.196	.167
5	.206	.271
6	.147	.240

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

placed at zero, were as follows: Word Generation, 16.3%; Sentence Generation, 9.6%; Phonological Decision, 12.1%; Word Decision, 32.2%; Sentence Decision, 14.7%; Syntactic Decision, 19.8%.

Evaluation of Prediction 4: On categorical analysis, individuals who depart from left-brained laterality on one or more tasks will be more likely to be left-handed than those who are consistently left-lateralised. Our preregistered analysis did not support this prediction, but this negative result should be interpreted cautiously.

It would be premature, given the observed data, to assume that there was no handedness effect on consistent left-lateralisation. First, the categorical analysis is less sensitive than the analysis of continuous laterality indices shown in Table 3, which clearly indicate reduced lateralisation for left-handers on a subset of tasks. Furthermore, as shown in our prior analyses, the different laterality indices are not independent, and hence Bonferroni correction, which assumes independence of measures, is over-conservative. We drew a binary divide at zero, but this means that many cases close to zero will not be clearly lateralised. An alternative approach would have been to make a three-way division between lateralised left, not lateralised, and lateralised right. The main reason for not doing that was that the numbers in the clearly lateralised groups would be relatively small, giving low power.

Nevertheless, although we cannot conclude there is no effect of handedness, it is evident that on the fTCD measures, the differences between left- and right-handers are modest, and smaller than reported in the literature on Wada testing. The most striking finding is that, when we use just a categorical left- versus right-hemisphere coding, a large proportion of people are not consistent in their direction of lateralisation across measures, regardless of handedness.

3.4. *Relationship between behavioural and fTCD laterality indices*

3.4.1. *Testing prediction 5: LIs will be similar for comparable behavioural and blood-flow measures*

Our fifth prediction was: ***the laterality profile obtained with the online language battery will be significantly associated with the profile seen with the direct measurement of cerebral blood flow using fTCD, with laterality on Dichotic Listening and Word Comprehension relating more strongly to receptive language tasks, and Rhyme Decision to language generation tasks.***

A preliminary inspection of correlations between online and fTCD laterality indices (Fig. 12) showed very little relationship between the two, even for the two measures, Word Comprehension and Rhyme Decision, that have analogues in fTCD (Word Decision and Phonological Decision respectively).

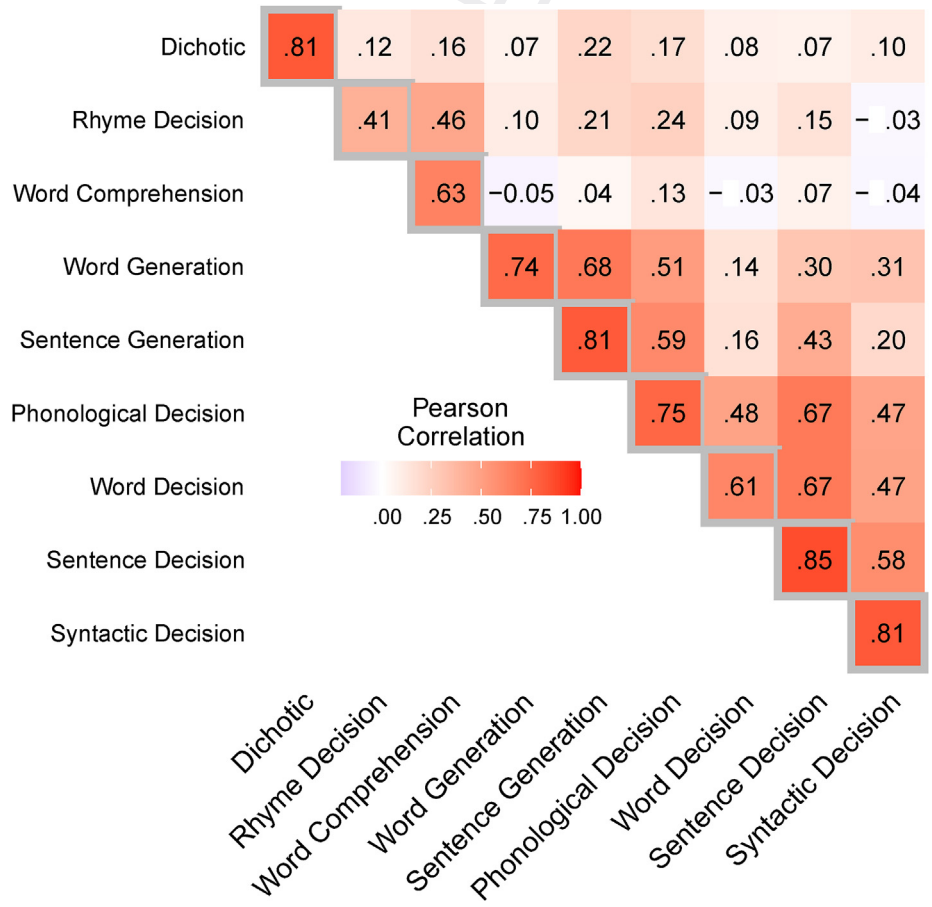

Fig. 12 – Heatmap with correlations between behavioural and fTCD laterality indices for both handedness groups combined. The diagonal (cells with grey frames) shows split-half reliability based on participants who completed both sets of measures.

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

We had preregistered two data checks: 1) Online measures that have split-half reliability below .6 will be excluded from further analysis. 2) Online measures of Word Comprehension and Rhyme Decision will only be taken forward to the next stage of analysis if they have a correlation of at least .11 with the counterpart measure from fTCD (Word Decision and Phonological Decision respectively).

The online Word Comprehension measure failed the second check: the correlation with the fTCD Word Decision laterality index was close to zero. The correlation between the online Rhyme Decision and fTCD Phonological Decision was .242, meeting our criterion, but split-half reliability was only .41 [.33, .48]. Accordingly, we proceeded with the next step of analysis only with Dichotic Listening (which had good reliability, but no counterpart in the fTCD battery).

We had predicted that Dichotic Listening, as a receptive task, should load on the same factor as the Word Decision, Sentence Decision and Syntactic Decision, but it is evident from the heatmap that, insofar as it correlates with the fTCD tasks, the strongest association is with Sentence Generation, a production task.

In practice, a model including a free path from Dichotic Listening to Factor 2 gave a better fit than a model with the path fixed to zero (Table 8): χ^2 difference = 12.99, DF = 1, $p < .001$. This is not a strong test of our prediction, because it will be passed if there is even a weak correlation between dichotic LI and fTCD laterality indices. Guided by the data, we ran an alternative model (not preregistered) with Dichotic Listening loading on Factor 1. This also gave excellent fit, with lower χ^2 than the preregistered model (for full results see Supplementary Material 8 (Available at: https://osf.io/g9tqh/)).

Evaluation of Prediction 5: Equivalence of behavioural and fTCD laterality indices

In our preregistration analysis plan, we stated: **If the online measures do not correspond to the parallel measures from fTCD this could mean that these measures are insufficiently reliable to index laterality in individuals – this would be evidenced by poor split-half reliability, and would indicate the need to either abandon this approach or to seek better measures. If the measures are reliable, but 95% confidence intervals for path estimates of online measures span zero, we can conclude that the online measures are tapping different aspects of laterality than the fTCD measures. If measures are reliable and good fit is obtained for a two-factor model that incorporates the laterality indices from online measures, this would support the use of online tests as proxy measures for underlying lateralised brain activation**

Overall, the associations between behavioural and fTCD laterality indices were low enough to give little confidence in the specific pattern of associations. Although the split-half reliability of the online measures was not high, it was not so low as to explain the lack of association with fTCD measures. The main conclusion is that our online behavioural measures of laterality based on speed of responding to lateralised stimuli have little in common with measures of relative blood flow to the two hemispheres while performing the same tasks, and cannot be used as proxy measures.

Table 8 – Fit statistics for models including Dichotic Listening.

Path to dichotic	DF	χ^2
No dichotic path	12	24.17
Dichotic <- Factor 1	11	11.18
Dichotic <- Factor 2	11	7.34

3.5. *Additional analysis of relationship between behavioural performance and laterality indices*

It seemed possible that the laterality indices might be influenced by either the language status of participants, and/or behavioural scores on the fTCD tasks (number of items generated or percentage correct). Supplementary file 6 (Available at: https://osf.io/g9tqh/) shows relevant analyses indicating that this is not the case.

4. Discussion

In this study, we measured individual differences in language laterality using two approaches: behavioural biases on online measures, and task-related blood flow to left and right hemispheres using fTCD. Our focus was on the extent to which laterality measures were associated, and whether the pattern of association differed for left- and right-handers.

We had preregistered five specific predictions, none of which was confirmed. Nevertheless, the study has taken forward our understanding of language laterality, by allowing us to dismiss certain hypotheses, suggesting new avenues for research, and evaluating comparability of different ways of measuring cerebral lateralisation.

To simplify the interpretation of this complex dataset, we focus first on two overarching questions addressed by the study. The first question concerns correlations between laterality measures: in brief, is there evidence for a single language laterality dimension on which people vary? The second question concerns handedness: does the answer to our first question differ in groups of left- and right-handers? In addition we consider specific issues arising in this dataset, namely the finding of right hemisphere lateralisation on Word Decision, and the lack of agreement between behavioural and fTCD laterality indices. Finally, we consider how the particular factor structure seen in the fTCD analysis might be explained.

4.1. *No support for a single laterality dimension*

Previous attempts to consider the dimensionality of language lateralisation have been obscured by two issues. First, many studies have been conducted with measures whose reliability was not established. If two laterality indices are not correlated, it could just be because they are unreliable, and so it has been easy to dismiss lack of correlation between laterality measures as uninformative. Second, researchers have tended to focus only on measures that show left-lateralisation at the population level, treating unlateralised language measures as uninteresting.

Considering first the online behavioural data, the most noteworthy observation was that the correlations between laterality indices from the three tasks were generally weak.

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

This could not be attributed solely to poor reliability: although reliabilities of the two new tasks, Rhyme Decision and Word Comprehension, were not impressive (ranging from .54 to .55 for test-retest), they were higher than the intercorrelations between measures. Our data was not suitable for a more formal model comparison, but inspection of the pattern of correlations between measures made it clear that our proposed two-factor model, with Dichotic Listening and Word Comprehension being positively correlated and unrelated to Rhyme Decision could not be supported. Indeed, the strongest correlation was found between Word Comprehension and Rhyme Decision, consistent with the idea that task demands (speeded responding to picture stimuli) might be a greater determinant of strength of lateralisation than whether receptive or expressive language was involved.

For the fTCD data, results were generally in good agreement with Woodhead et al. (2021), with good reliability of LIs on most tasks, even though some tasks were not left-lateralised. As in our previous study (Woodhead et al., 2019), the LI on Syntactic Decision task had good reliability, and a distinctive pattern of association with other LIs, despite being unlateralised. This observation shows that lack of lateralisation at the population level does not mean that all individuals use both hemispheres equally for the task: it seems rather that the population contains a mixture of people, some of whom consistently prefer the left hemisphere, others the right, and others more equally balanced. The LI from the Word Decision task was the least reliable in the battery, yet again showed quite distinctive patterns of selective association with other tasks.

With fTCD we were able to subject the single factor model to a stronger test, because we had sufficient tasks for Structural Equation Modeling. Consistent with Woodhead et al. (2021), we could reject a single factor model; this gave a poor fit to the data, as it could not account for the fact that the correlations between LIs tended to form clusters. We tested a preregistered, alternative two-factor model that involved a division between language generation and language reception. This accounted for significantly more variance than the single-factor model, but still left a great deal unexplained, and overall the fit was poor. Accordingly, following our preregistration, we divided the sample into two sub samples to explore different models and found one that gave good fit, which was then replicated in the second half of the sample. This again had a two factor structure, but had two of the tasks, Phonological Decision and Syntactic Comprehension, loading on both the factors.

In a final step of analysis, we considered adding the laterality indices from online tasks to the model. The correlations between LIs from online tasks and fTCD were generally weak, and these measures did not help differentiate models. A model that included Dichotic Listening gave better fit when a non-zero path was included, than when it was set to zero, but the best fit was seen for a model where Dichotic Listening loaded on Factor 1 (with language generation tasks), rather than for our prespecified model where Dichotic Listening was regarded as an indicator of Factor 2 (with receptive tasks).

4.2. *Left- versus right-handers*

For both behavioural and fTCD LIs, with just one exception, there was a consistent trend for stronger left-hemisphere bias in right-handers than in left-handers. This reached significance on all measures except fTCD Word Decision. The exception was online Rhyme Judgement, which was not left-lateralised and where means for left- and right-handers were very similar.

The SEM analysis allowed us to go beyond simple comparison of means to test whether the association between LIs showed a similar pattern in the two handedness groups. In our previous fTCD study using four of the same measures (Woodhead et al., 2021), we had concluded that there was more dissociation between factors in left-handers, but this finding was not replicated here. There are two possible explanations for this discrepancy. First, the sample size in both the current and the previous study was small for this kind of analysis, raising the possibility that the initial finding was a false positive, or the current finding a false negative. Even with the current sample size, power to detect model invariance is not high. Second, the language tasks used with fTCD differed across studies. In the Woodhead et al. study, we included a List Generation task, that was intended to be a relatively pure measure of phonological output, using overlearned sequences, as well as a Semantic Decision task, which involved judging whether two pictures were semantically related. The List Generation task was dropped because of low reliability, and the Word Decision task was substituted for Semantic Decision to give a purer measure of comprehension at the single word level. We cannot rule out the possibility that these tests would show different patterns of association in left- and right-handers. Note, however, that the data did not support an additional possibility, namely that left-handers are more variable than right-handers. The comparison of factor scores in the two handedness groups showed differences in means, but only small differences in standard deviations.

Overall, we can conclude that the current data suggest that there is no reason to postulate different models for left- versus right-handers. The substantial differences between these groups could be entirely accounted for in terms of differences in factor means, and was driven primarily by differences on Factor 1, which related to tasks involving language generation.

These quantitative differences between handedness groups did not, however, translate into pronounced differences in proportions who were atypically lateralised on individual tasks, and most people had a mixture of left- and right-lateralisation across the whole fTCD battery.

4.3. *Right hemisphere lateralisation for comprehension of single words*

The finding of a slight bias to right hemisphere lateralisation for the behavioural Word Comprehension task and Word Decision on fTCD. Woodhead et al. (2021), using fTCD, showed left lateralisation in right-handers for a semantic decision task that involved judging if two (unnamed) pictured items were semantically related. A meta-analysis of fMRI studies by Vigneau et al. (2011) found that right-hemisphere involvement in lexical-semantic tasks was extremely limited. However, tasks used in fMRI typically involve thinking explicitly about meaning; e.g., Binder et al. (1996) devised a classic semantic decision task that is highly left-lateralised, which involves listening to animal names and deciding if they met specific

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

semantic criteria (e.g., "native to the United States"). Another consideration is that written, rather than auditory, presentation is commonly used in fMRI studies to avoid interference from scanner noise. In contrast, our current task required only a direct matching of a spoken word to one of a pair of semantically-related pictures. This task was the easiest in the battery, with near-ceiling performance by all participants, and it could be that it was insufficiently challenging to engage lateralised language systems. It might be tempting to suppose that right-hemisphere bias was induced by the left-right scanning of pictured items, but if that were the case, we should also have seen this bias in Rhyme Decision (which was weakly left-lateralised), and we should not have found a rightward bias in the fTCD Word Decision task, where pictures are presented vertically.

The relatively low reliability of the task, in both behavioural and fTCD formats, indicates the need for caution in interpretation. We might be tempted to dismiss the result, except for the fact that the bias was found both in the online behavioural version of the task (see Table 2 and in the analogous Word Decision fTCD task (see Table 3). Furthermore, on the behavioural task, left-handers showed a stronger effect than right-handers, in line with their reduced left-hemisphere bias on other tasks. This consistent picture, however, is challenged by the fact that the correlation between LIs on the online and fTCD versions of the task was close to zero.

Overall, the pattern of results is puzzling, and we will need more reliable measures of single word comprehension to determine whether it is meaningful. It does raise the possibility that where semantic decision tasks are lateralised, this may relate to task demands that involve explicit reasoning about word meanings.In addition, we need to be aware that other, nonlinguistic, factors, such as presentation rate may also affect laterality indices (Payne, Gutierrez-Sigut, Subik, Woll, & MacSweeney, 2015).

4.4. *Differences between behavioural and fTCD measures of laterality*

The lack of agreement between online and fTCD laterality indices is disappointing for those hoping to use behavioural measures as proxies for more direct brain measures of laterality. It is, of course, possible that stronger associations might be seen with better measures: behavioural measures obtained under laboratory conditions are likely to be more reliable than those obtained online; the tasks we developed for fTCD might have limitations in terms of variation in strategies used by participants, or in the impact of nonlinguistic task demands that could also engage lateralised systems. Nevertheless, even for Dichotic Listening, a task that has a long track record as a behavioural laterality index, and Word Generation, the gold standard laterality measure in fTCD, the correlation between LIs was weak. This is despite the fact that both tasks are consistently lateralised with good reliability. Furthermore, validity of the online Dichotic Listening task is supported by the fact that it showed a small but reliable difference in laterality for left- and right-handers, similar to that previously reported by Karlsson, Johnstone, and Carey (2019), who used in-person rather than online testing.

Several researchers have proposed that when optimal methods are used, dichotic listening and/or visual half-field methods can be useful indicators of language laterality (Van der Haegen, Westerhausen, Hugdahl, & Brysbaert, 2013; Westerhausen, 2019), and it would seem that, when the goal is to categorise individuals as left- or right-lateralised for language, strong lateralisation on such behavioural measures is usually a good indicator of laterality on brain measures. Nevertheless, many people do not show such strong lateralisation, and when studies have considered quantitative associations between laterality indices from dichotic listening and fMRI, correlations have been unimpressive; e.g., Bethmann et al. (2007) reported a correlation of .38, finding that many individuals with a left-ear advantage on dichotic listening were left-lateralised for language on a synonym decision task. Other studies have found higher correlations than those reported in the current study, but the level of agreement is moderate at best, and small samples mean that there will be large confidence intervals around these estimates. Data extracted from a scatterplot by Van der Haegen et al. (2013) gave a Spearman correlation between dichotic listening and fMRI (word generation) of .46 (N = 62), in a sample that excluded cases of ambiguous laterality on fMRI (those with absolute value of LI below .6). Hund-Georgiadis, Lex, Friederici, and von Cramon (2002) tested 17 left-handers and 17 right-handers and found Spearman correlation of .56 between laterality indices from dichotic listening and semantic encoding on fMRI (computed from data extracted from scatterplot). Gerrits, De Clercq, Verhelst, and Vingerhoets (2020) found that a laterality index from a visual half-field naming task gave a Spearman correlation of .54, (95% CI .31, .71) with laterality on word generation from fMRI in a sample of 63 left-handers, 38 of whom were selected as candidates for right-hemisphere language because they had a difference of at least 20 msec in response time in favour of the left versus right visual field. On categorical assignment based on fMRI, 58% of the candidate cases had right-hemisphere language confirmed. This indicates that on the one hand, behavioural tasks have potential to screen for cases of atypical lateralisation, but on the other hand, agreement between behavioural laterality and fMRI laterality is imperfect, even when the behavioural task involves speech production.

It is well worth pursuing the goal of standardizing behavioural laterality measures and working towards optimising their reliability, as exemplified by work by Westerhausen and Samuelsen (2020). We suspect, however, that the lack of agreement between different kinds of laterality measure might not be resolved simply by improving reliability. If, as we propose, language laterality is not unitary, then the agreement between different measures will depend on the relative contribution of different lateralised systems. In this regard, it is worth noting that fMRI studies of dichotic listening indicate activation that extends well beyond the temporal lobes to

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

include bilateral activity in frontal lobes, which seems in part dependent on the inhibitory demands of the task (Jäncke & Shah, 2002; Westerhausen, Kompus, & Hugdahl, 2014).

4.5. *Interpreting the two-factor structure*

Although our data did not support the simple two-factor structure of laterality that we predicted, we were able to obtain a good fit for a model that included additional paths. This two-factor account is reminiscent of the dual stream model of Hickok and Poeppel (2007), who postulated a dorsal stream from superior temporal to premotor cortices via the arcuate fasciculus, and a ventral stream from temporal cortex to anterior inferior frontal gyrus. The former is implicated in integrating auditory speech with articulator motor actions, and is lateralised, whereas the latter is not lateralised, and is involved in access to conceptual memory and mapping of sound to meaning. However, as noted in our previous study, the exclusive loading of Sentence Generation on factor 1, and the loading of Sentence Decision on both factors is not entirely consistent with the dual stream account. We also previously found List Generation, which would be expected to involve the dorsal stream, was not lateralised.

We conclude by considering what commonalities there are between tasks that characterise each factor.

Table 9 summarises the characteristics of the six fTCD tasks, grouped according to the factors they load on. The online Dichotic Listening task is also shown. The task battery had been designed to include three tasks that involved language generation, and three that involved receptive language. Note that, in contrast to most tasks used in fMRI studies, spoken language was used to present stimuli for the receptive tasks. For Syntactic Decision this was supplemented with written words, as the task was otherwise too difficult. As predicted, the language generation tasks loaded on Factor 1 and the receptive tasks on Factor 2, but in addition Phonological Decision loaded on Factor 2, and Sentence Decision on Factor 1. Phonological Decision, unlike the other tasks loading on Factor 2, did not involve auditory input, but did have in common the 2-choice response format of other Factor 2 tasks.

Sentence Decision also behaved unexpectedly, in that it had significant loadings on Factor 1, despite being designed as a purely receptive task. Unlike the other two receptive tasks, Sentence Decision requires the listener to use syntactic information to assign semantic roles and build meaning representations, and hence more linguistic computation than the receptive tasks that used single word stimuli (Word Decision) or meaningless material (Syntactic Decision).

As seen in Fig. 7, the four tasks that load on Factor 1 are all significantly lateralised, at least in right-handers. In contrast, those loading on Factor 2 include two tasks that are not significantly lateralised.

In interpreting this finding, we should first rule out two trivial explanations for the factor structure uncovered in SEM. First, this structure cannot be regarded as an artefact of including tasks differing in degree of laterality. This is because factor structure in SEM is computed solely on the basis of covariances between measures, and means do not affect it. Thus, we could add a constant to the LIs for tasks D and E to make them lateralised, and the factor solution would remain the same.

Second, it is unlikely that the pattern of results is simply due to contamination of the laterality index by non-linguistic activations. This is because with fTCD we do a direct subtraction of blood flow velocity in left and right hemispheres. Activation due, for instance, to visual processing, would influence the laterality index only if such activation were also lateralised.

Third, we can rule out an explanation that treats the non-lateralised tasks as not relevant for studying individual differences in laterality. Such an explanation would be justified if tasks such as Word Decision, Sentence Decision and Syntactic Decision were simply unreliable. Low reliability would be expected if these tasks were not lateralised in individuals, because people used both hemispheres jointly, or switched from one to the other at random. The new task, Word Decision, was the least reliable in the battery, but nevertheless, the split-half reliability was moderate. The other two receptive tasks had good test-retest reliability in our previous study (Woodhead et al., 2021) and good split-half reliability in the current study. Thus, even though there is weak or absent lateralisation at the population level on these tasks, the degree and direction of lateralisation is reasonably consistent within individuals. And indeed, if that were not the case, we

Table 9 – Characteristics of tasks.

Task	Input modality	Output	Active Passive	Language component	Factor
Word generation	visual (letters)	speech	active generation	Phonology, lexicon	1
Sentence generation	visual (complex scene)	speech	active generation	Phonology, syntax, semantics	1
Phonological/ Rhyme decision	visual (2 pictures)	yes/no keypress	generate names, + 2-choice decision	Phonology, lexicon	1 + 2
Sentence decision	visual (2 complex scenes) +auditory (spoken sentence)	L/R keypress	2-choice decision	Phonology, syntax, semantics	1 + 2
Word decision/ comprehension	visual (2 pictures) + auditory (spoken word)	L/R keypress	2-choice decision	Phonology, lexicon	2
Syntactic decision	visual (written nonwords) + auditory (spoken nonwords)	yes/no keypress	2-choice decision	Phonology, syntax	2
Dichotic listening	auditory (syllables), competing	keypress	6-choice decision	Phonology	[1]

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

would not expect the tasks to show moderate intercorrelations with one another.

To account for the observed pattern of results, we postulate two language centres, one lateralised at the population level (centre L), and the other centred on zero (centre Z). An individual's observed fTCD laterality on a task will depend on the extent to which these two centres are implicated in task performance, with Word Generation and Sentence Generation being largely dominated by centre L, Syntactic Decision and Word Decision by centre Z, and Phonological Decision and Sentence Decision implicating both centres.

To some extent, this is less of an explanation than a redescription of the data, but it does yield novel predictions that can be tested using fMRI, which gives information on localisation of activation within a hemisphere. The prediction would be that there would be more overlap in brain regions activated by tasks that load on the same factor than for those loading on different factors, and furthermore, activation would be lateralised only for brain regions supporting Factor 1 tasks. The BIL&GIN fMRI study (Mazoyer et al., 2016) included data from tasks analogous to those used here and could be used to test these predictions.

5. Conclusions

Language laterality in individuals is not a single dimension, but varies depending on task demands. A simple two-factor model that distinguished tasks involving language generation and comprehension was superior to a single-factor model in accounting for individual variation, but tasks did not neatly map onto the two factors. Previously, we had suggested that a different factor structure would be seen in left- and right-handers, but that was not the case for the current dataset: handedness determined the mean level of laterality on a language generation factor, but not the covariance between factors. Behavioural laterality measures, assessed online, were only weakly related to laterality measured using fTCD. The full dataset provides unique possibilities for assessing associations between different laterality measures and is openly available for exploration by other researchers without restriction.

Author contributions

Adam J. Parker: Conceptualisation, Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing – Review and Editing, Visualisation, Supervision; Zoe V. J. Woodhead: Conceptualisation, Methodology, Validation, Formal analysis, Investigation, Data Curation, Writing – Review and Editing, Visualisation, Supervision; David P. Carey: Conceptualisation, Methodology, Writing – Review and Editing, Supervision; Margriet A. Groen: Conceptualisation, Writing – Review and Editing, Supervision; Eva Gutierrez-Sigut: Methodology, Writing – Review and Editing; Jessica Hodgson: Methodology, Writing – Review and Editing, Supervision; John Hudson: Supervision; Emma M. Karlsson: Conceptualisation, Methodology, Writing – Review and Editing; Mairéad MacSweeney: Conceptualisation, Methodology, Writing – Review and Editing, Supervision; Heather Payne: Methodology, Writing – Review and Editing; Nuala Simpson: Project Administration; Paul A. Thompson: Formal analysis, Data Curation; Writing – Review and Editing; Kate E. Watkins: Conceptualisation, Writing – Review and Editing, Ciara Egan: Methodology, Investigation; Jack H. Grant: Methodology, Investigation; Sophie Harte: Methodology, Investigation; Brad T. Hudson: Methodology, Investigation; Maria Sablik: Investigation; Nicholas A. Badcock: Investigation, Writing – Review and Editing; Dorothy V. M. Bishop: Conceptualisation, Methodology, Formal analysis, Writing – original draft, Supervision, Funding Acquisition.

Data availability

Raw data, analysis scripts and materials are all available on Open Science Framework: https://osf.io/g9tqh/.

Open practices

The study in this article earned Open Data, Open Materials and Preregistered badges for transparent practices. Materials and data for the study are available at https://osf.io/g9tqh/.

Uncited references

Q6

Balota et al., 2007; Bryden, 1975; Champely, 2018; Fesl et al., 2010; Fletcher, 2010; Moshagen, 2020; Moshagen and Erdfelder, 2016; Ocklenburg et al., 2016; Pornprasertmanit et al., 2020; Taylor et al., 2020; Wagenmakers and Farrell, 2004; Wang and Wang, 2012; Wolf et al., 2013.

Declaration of competing interest

The authors declare no competing interests.

Acknowledgements

Q5 This study is supported by an Advanced Grant from the European Research Council [694189]. We thank Melissa Gibbs, Hussein Mehmet, Nahian Tasnim Nur and Louis Pitsikas for assistance with recruitment and data collection.

REFERENCES

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, 52(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, 39(3), 445–459.

Bethmann, A., Tempelmann, C., De Bleser, R., Scheich, H., & Brechmann, A. (2007). Determining language laterality by fMRI

and dichotic listening. *Brain Research, 1133*, 145–157. https://doi.org/10.1016/j.brainres.2006.11.057

Binder, J. R., Swanson, S. J., Hammeke, T. A., Morris, G. L., Mueller, W. M., & Fischer, M. (1996). Determination of language dominance using functional MRI: A comparison with the Wada test. *Neurology, 46*(4), 978–984. https://doi.org/10.1212/wnl.46.4.978

Bishop, D. V. M. (1980). measuring familial sinistrality. *Cortex, 16*, 311–313.

Bishop, D. V. M. (1990). On the futility of using familial sinistrality to subclassify handedness groups. *Cortex, 26*, 153–155.

Bishop, D. V. M. (2003). *The test for reception of grammar, version 2 (TROG-2)*. London: Pearson.

Bishop, D. V. M. (2013). Cerebral asymmetry and language development: Cause, correlate, or consequence? *Science, 340*(6138). https://doi.org/10.1126/science.1230531

Bless, J. J., Westerhausen, R., Arciuli, J., Kompus, K., Gudmundsen, M., & Hugdahl, K. (2013). Right on all occasions?"– on the feasibility of laterality research using a smartphone dichotic listening application. *Frontiers in Psychology, 4*, 42.

Bradshaw, A. R., Bishop, D. V., & Woodhead, Z. V. (2017). Methodological considerations in assessment of language lateralisation with fMRI: A systematic review. *PeerJ, 5*, e3557. https://doi.org/10.7717/peerj.3557

Bryden, M. P. (1975). Speech lateralisation in families: A preliminary study using dichotic listening. *Brain and Language, 2*, 201–211.

Bryden, M. P. (1982). *Laterality: Functional asymmetry in the intact brain*. New York: Academic Press.

Carey, D. P., & Johnstone, L. T. (2014). Quantifying cerebral asymmetries for language in dextrals and adextrals with random-effects meta analysis. *Frontiers in Psychology, 5*, 1128. https://doi.org/10.3389/fpsyg.2014.01128

Champely, S. (2018). *pwr: Basic functions for power analysis*. R package version 1.2-2. Retrieved from https://CRAN.R-project.org/package=pwr.

Corey, D., & Foundas, A. (2010). Measuring familial sinistrality: Problems with dichotomous classification. *Laterality, 10*(4), 321–335. https://doi.org/10.1080/13576500442000111

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *The Quarterly Journal of Experimental Psychology: QJEP, 71*(4), 808–816. https://doi.org/10.1080/17470218.2017.1310261

Fedorenko, E., Hsieh, P.-J., Nieto-Castañón, A., Whitfield-Gabrieli, S., & Kanwisher, N. (2010). New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology, 104*(2), 1177–1194. https://doi.org/10.1152/jn.00032.2010

Fesl, G., Bruhns, P., Rau, S., Wiesmann, M., Ilmberger, J., Kegel, G., … P. M.. (2010). Sensitivity and reliability of language laterality assessment with a free reversed association task-a fMRI study. *European Radiology, 20*(3), 683–695. https://doi.org/10.1007/s00330-009-1602-4

Fletcher, T. D. (2010). *Psychometric: Applied psychometric theory*. R package version 2.2. Retrieved from https://CRAN.R-project.org/package=psychometric.

Gerrits, R., De Clercq, P., Verhelst, H., & Vingerhoets, G. (2020a). Evaluating the performance of the visual half field paradigm as a screening tool to detect right hemispheric language dominance. *Laterality, 25*(6), 722–739. https://doi.org/10.1080/1357650X.2020.1854279

Gerrits, R., Verhelst, H., & Vingerhoets, G. (2020b). Mirrored brain organization: Statistical anomaly or reversal of hemispheric functional segregation bias? *Proceedings of the National Academy of Sciences, 117*, 14057–14065. https://doi.org/10.1073/pnas.2002981117

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition, 177*, 263–277. https://doi.org/10.1016/j.cognition.2018.04.007

Hervé, M. (2021). *RVAideMemoire: Testing and plotting procedures for biostatistics*. R package version 0.9-79.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews. Neuroscience, 8*(5), 393–402. https://doi.org/10.1038/nrn2113

Hoaglin, D. C., & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labelling. *The Journal of the Acoustical Society of America, 82*(400), 1147–1149.

Hugdahl, K., & Andersson, L. (1986). The "forced-attention paradigm" in dichotic listening to CV-syllables: A comparison between adults and children. *Cortex, 22*(3), 417–432.

Hund-Georgiadis, M., Lex, U., Friederici, A. D., & von Cramon, D. Y. (2002). Non-invasive regime for language lateralization in right- and left-handers by means of functional MRI and dichotic listening. *Experimental Brain Research, 145*(2), 166–176. https://doi.org/10.1007/s00221-002-1090-0

Hunter, Z. R., & Brysbaert, M. (2008). Visual half-field experiments are a good measure of cerebral language dominance if used properly: Evidence from fMRI. *Neuropsychologia, 46*(1), 316–325.

Jäncke, L., & Shah, N. J. (2002). Does dichotic listening probe temporal lobe functions? *Neurology, 58*(5), 736–743. https://doi.org/10.1212/wnl.58.5.736

Johnstone, L. T., Karlsson, E. M., & Carey, D. P. (2020). The validity and reliability of quantifying hemispheric specialisation using fMRI: Evidence from left and right handers on three different cerebral asymmetries. *Neuropsychologia, 138*, Article 107331. https://doi.org/10.1016/j.neuropsychologia.2020.107331

Karlsson, E. M., Johnstone, L. T., & Carey, D. P. (2019). The depth and breadth of multiple perceptual asymmetries in right handers and non-right handers. *Laterality, 24*(6), 707–739. https://doi.org/10.1080/1357650X.2019.1652308

Kline, R. B. (2011). *Principals and practice of structural equation modeling* (3rd ed.). Guilford Press.

Knecht, S., Deppe, M., Dräger, B., Bobe, L., Lohmann, H., Ringelstein, E.-B., & Henningsen, H. (2000). Language lateralization in healthy right-handers. *Brain, 123*(1), 74–81. https://doi.org/10.1093/brain/123.1.74

Knecht, S., Deppe, M., Ebner, A., Henningsen, H., Huber, T., Jokeit, H., & Ringelstein, E. B. (1998). Noninvasive determination of language lateralization by functional transcranial Doppler sonography A comparison with the Wada test. *Stroke, 29*(1), 82–86. https://doi.org/10.1161/01.STR.29.1.82

Lee, D., Swanson, S. J., Sabsevitz, D. S., Hammeke, T. A., Scott Winstanley, F., Possing, E. T., & Binder, J. R. (2008). Functional MRI and Wada studies in patients with interhemispheric dissociation of language functions. *Epilepsy & Behavior: E&B, 13*(2), 350–356. https://doi.org/10.1016/j.yebeh.2008.04.010

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced Learners of English. *Behavior Research Methods, 44*(2), 325–343. https://doi.org/10.3758/s13428-011-0146-0

Mazoyer, B., Mellet, E., Perchey, G., Zago, L., Crivello, F., Jobard, G., Delcroix, N., Vigneau, M., Leroux, G., Petit, L., Joliot, M., & Tzourio-Mazoyer, N. (2016). BIL&GIN: A neuroimaging, cognitive, behavioral, and genetic database for the study of human brain lateralization. *Neuroimage, 124*, 1225–1231. https://doi.org/10.1016/j.neuroimage.2015.02.071

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

Mazoyer, B., Zago, L., Jobard, G., Crivello, F., Joliot, M., Perchey, G., … Tzourio-Mazoyer, N. (2014). Gaussian Mixture Modeling of hemispheric lateralization for language in a large sample of healthy individuals balanced for handedness. *Plos One, 9*, Article e101165. https://doi.org/10.1371/journal.pone.0101165

McKeever, W. F., & Vandeventer, A. D. (1977). Visual and auditory language processing asymmetries: Influences of handedness, familial sinistrality, and sex. *Cortex, 13*, 225–241.

Moshagen, M. (2020). *semPower: Power analyses for SEM*. R package version 1.0.1. Retrieved from https://CRAN.R-project.org/package=semPower.

Moshagen, M., & Erdfelder, E. (2016). A new strategy for testing structural equation models. *Structural Equation Modeling, 23*, 54–60. https://doi.org/10.1080/10705511.2014.950896

Ocklenburg, S., Ströckens, F., Bless, J. J., Hugdahl, K., Westerhausen, R., & Manns, M. (2016). Investigating heritability of laterality and cognitive control in speech perception. *Brain and Cognition, 109*, 34–39. https://doi.org/10.1016/j.bandc.2016.09.003

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia, 9*, 97–113.

Orsini, D. L., Satz, P., Soper, H. V., & Light, R. K. (1985). The role of familial sinistrality in cerebral organization. *Neuropsychologia, 23*(2), 223–232. https://doi.org/10.1016/0028-3932(85)90106-x

Parker, A. J., Woodhead, Z. V. J., Thompson, P. A., & Bishop, D. V. M. (2021). Assessing the reliability of an online behavioural laterality battery: A pre-registered study. *Laterality, 26*(4), 359–397. https://doi.org/10.1080/1357650X.2020.1859526

Payne, H., Gutierrez-Sigut, E., Subik, J., Woll, B., & MacSweeney, M. (2015). Stimulus rate increases lateralisation in linguistic and non-linguistic tasks measured by functional transcranial Doppler sonography. *Neuropsychologia, 72*, 59–69. https://doi.org/10.1016/j.neuropsychologia.2015.04.019 .1080/1357650X.2020.1859526

Phillips, N. (2017). *Yarrr: A companion to the e-book "YaRrr!: The pirate's guide to R"*. R package version 0.1.5. https://CRAN.R-project.org/package=yarrr.

Porac, C., & Coren, S. (1976). The dominant eye. *Psychological Bulletin, 83*(5), 880–897. https://doi.org/10.1037/0033-2909.83.5.880

Pornprasertmanit, S., Miller, P., Schoemann, A., & Jorgensen, T. D. (2020). *simsem: SIMulated structural equation modeling*. R package version 0.5-15. https://CRAN.R-project.org/package=simsem.

R Core Team. (2016). *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Ramsey, N. F., Sommer, I. E. C., Rutten, G. J., & Kahn, R. S. (2001). Combined analysis of language tasks in fMRI improves assessment of hemispheric dominance for language functions in individual subjects. *Neuroimage, 13*(4), 719–733. https://doi.org/10.1006/nimg.2000.0722

Rasmussen, T., & Milner, B. (1975). Clinical and surgical studies of the cerebral speech areas in man. In K. Zülch, O. Creutzfeldt, & G. Galbraith (Eds.), *Cerebral localisation* (pp. 238–257). New York: Springer-Verlag.

Rasmussen, T., & Milner, B. (1977). The role of early left-brain injury in determining lateralization of cerebral speech functions. *Annals of the New York Academy of Sciences, 299*(1), 355–369. https://doi.org/10.1111/j.1749-6632.1977.tb41921.x

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*, 1–36.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Tests of significance and goodness-of-fit measures. *Methods of Psychological Research, 8*(2), 23–74.

Seghier, M. L. (2008). Laterality index in functional MRI: Methodological issues. *Magnetic Resonance Imaging, 26*(5), 594–601. https://doi.org/10.1016/j.mri.2007.10.010

Sørensen, Ø., & Westerhausen, R. (2020). From observed laterality to latent hemispheric differences: Revisiting the inference problem. *Laterality, 25*(5), 560–582. https://doi.org/10.1080/1357650X.2020.1769124

Taylor, J. E., Beith, A., & Sereno, S. C. (2020). LexOPS: An R package and user interface for the controlled generation of word stimuli. *Behavior Research Methods*. https://doi.org/10.3758/s13428-020-01389-1

Van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67.

Van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *The Quarterly Journal of Experimental Psychology: QJEP, 67*, 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Van der Haegen, L., & Brysbaert, M. (2018). The relationship between behavioral language laterality, face laterality and language performance in left-handers. *Plos One, 13*(12), 1–22. https://doi.org/10.1371/journal.pone.0208696

Van der Haegen, L., Westerhausen, R., Hugdahl, K., & Brysbaert, M. (2013). Speech dominance is a better predictor of functional brain asymmetry than handedness: A combined fMRI word generation and behavioral dichotic listening study. *Neuropsychologia, 51*(1), 91–97. https://doi.org/10.1016/j.neuropsychologia.2012.11.002

Vigneau, M., Beaucousin, V., Hervé, P.-Y., Jobard, G., Petit, L., & Crivello, F. (2011). What is right-hemisphere contribution to phonological, lexico-semantic, and sentence processing? Insights from a meta-analysis. *Neuroimage, 54*(1), 577–593. https://doi.org/10.1016/j.neuroimage.2010.07.036

Vingerhoets, G. (2019). Phenotypes in hemispheric functional segregation? Perspectives and challenges. *Physics of Life Reviews, 30*, 1–18. https://doi.org/10.1016/j.plrev.2019.06.002

Voyer, D. (1998). On the reliability and validity of noninvasive laterality measures. *Brain and Cognition, 36*, 209–236.

Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*(1), 192–196.

Wang, J., & Wang, X. (2012). *Structural equation modeling: Applications using mplus*. Chichester, West Sussex, U.K: Wiley/Higher Education Press.

Westerhausen, R. (2019). A primer on dichotic listening as a paradigm for the assessment of hemispheric asymmetry. *Laterality: Asymmetries of Body, Brain and Cognition, 24*(6), 740–771. https://doi.org/10.1080/1357650X.2019.1598426

Westerhausen, R., Kompus, K., & Hugdahl, K. (2014). Mapping hemispheric symmetries, relative asymmetries, and absolute asymmetries underlying the auditory laterality effect. *Neuroimage, 84*, 962–970. https://doi.org/10.1016/j.neuroimage.2013.09.074

Westerhausen, R., & Samuelsen, F. (2020). An optimal dichotic-listening paradigm for the assessment of hemispheric dominance for speech processing. *Plos One, 15*(6), Article e0234665. https://doi.org/10.1371/journal.pone.0234665

Wilke, M., & Lidzba, K. (2007). LI-tool: A new toolbox to assess lateralization in functional MR-data. *Journal of Neuroscience Methods, 163*(1), 128–136. https://doi.org/10.1016/j.jneumeth.2007.01.026

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for Structural Equation Models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*(6), 913–934.

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013

Woodhead, Z. V. J., Bradshaw, A. R., Wilson, A. C., Thompson, P. A., & Bishop, D. V. M. (2019). Testing the unitary theory of language lateralization using functional transcranial Doppler sonography in adults. *Royal Society Open Science, 6*(3), Article 181801. https://doi.org/10.1098/rsos.181801

Woodhead, Z. V. J., Rutherford, H., & Bishop, D. V. M. (2018). Measurement of language laterality using functional transcranial Doppler ultrasound: A comparison of different tasks [version 2; peer review: 2 approved]. *Wellcome Open Research, 3*, 104. https://doi.org/10.12688/wellcomeopenres.14720.2

Woodhead, Z. V. J., Thompson, P. A., Karlsson, E. M., & Bishop, D. V. M. (2021). An updated investigation of the multidimensional structure of language lateralization in left- and right-handed adults: A test–retest functional transcranial Doppler sonography study with six language tasks. *Royal Society Open Science, 8*(2), Article 200696. https://doi.org/10.1098/rsos.200696

Please cite this article as: Parker, A. J et al., Inconsistent language lateralisation – Testing the dissociable language laterality hypothesis using behaviour and lateralised cerebral blood flow, Cortex, https://doi.org/10.1016/j.cortex.2022.05.013